THE OXFORD HANDBOOK OF

PROBABILITY AND PHILOSOPHY

THE OXFORD HANDBOOK OF

PROBABILITY AND PHILOSOPHY

Edited by

ALAN HÁJEK

and

CHRISTOPHER HITCHCOCK

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© in this volume the several contributors 2016

The moral rights of the authors have been asserted

First Edition published in 2016

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data

Data available

Library of Congress Control Number: 2016933498

ISBN 978–0–19–960761–7

Printed in Great Britain by
Clays Ltd, St Ives plc

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

Contents

List of Contributors ix

Introduction 1
Alan Hájek and Christopher Hitchcock

1. Probability for Everyone—Even Philosophers 5
Alan Hájek and Christopher Hitchcock

PART I: HISTORY

2. Pre-history of Probability 33
James Franklin

3. Probability in 17th- and 18th-century Continental Europe from the Perspective of Jacob Bernoulli's *Art of Conjecturing* 50
Edith Dudley Sylla

4. Probability and Its Application in Britain during the 17th and 18th Centuries 69
David R. Bellhouse

5. A Brief History of Probability Theory from 1810 to 1940 87
Hans Fischer

6. The Origins of Modern Statistics: The English Statistical School 112
John Aldrich

7. The Origins of Probabilistic Epistemology: Some Leading 20th-century Philosophers of Probability 130
Maria Carla Galavotti

PART II: FORMALISM

8. Kolmogorov's Axiomatization and Its Discontents 155
Aidan Lyon

9. Conditional Probability 167
Kenny Easwaran

10. The Bayesian Network Story 183
Richard E. Neapolitan and Xia Jiang

PART III: ALTERNATIVES TO STANDARD PROBABILITY THEORY

11. Mathematical Alternatives to Standard Probability that Provide Selectable Degrees of Precision 203
Terrence L. Fine

12. Probability and Nonclassical Logic 248
J. Robert G. Williams

13. A Logic of Comparative Support: Qualitative Conditional Probability Relations Representable by Popper Functions 277
James Hawthorne

14. Imprecise and Indeterminate Probabilities 296
Fabio G. Cozman

PART IV: INTERPRETATIONS AND INTERPRETIVE ISSUES

15. Symmetry Arguments in Probability 315
Sandy Zabell

16. Frequentism 341
Adam La Caze

17. Subjectivism 360
Lyle Zynda

18. Bayesianism vs. Frequentism in Statistical Inference 382
Jan Sprenger

19. The Propensity Interpretation 406
Donald Gillies

20. Best System Approaches to Chance 423
Wolfgang Schwarz

21. Probability and Randomness 440
Antony Eagle

22. Chance and Determinism 460
Roman Frigg

PART V: PROBABILISTIC JUDGMENT AND ITS APPLICATIONS

23. Human Understandings of Probability 477
MICHAEL SMITHSON

24. Probability Elicitation 497
STEPHEN C. HORA

25. Probabilistic Opinion Pooling 519
FRANZ DIETRICH AND CHRISTIAN LIST

PART VI: APPLICATIONS OF PROBABILITY: SCIENCE

26. Quantum Probability: An Introduction 545
GUIDO BACCIAGALUPPI

27. Probabilities in Statistical Mechanics 573
WAYNE C. MYRVOLD

28. Probability in Biology: The Case of Fitness 601
ROBERTA L. MILLSTEIN

PART VII: APPLICATIONS OF PROBABILITY: PHILOSOPHY

29. Probability in Epistemology 625
MATTHEW KOTZEN

30. Confirmation Theory 650
VINCENZO CRUPI AND KATYA TENTORI

31. Self-Locating Credences 666
MICHAEL G. TITELBAUM

32. Probability in Logic 681
HANNES LEITGEB

33. Probability in Ethics 705
DAVID MCCARTHY

34. Probability and the Philosophy of Religion 738
PAUL BARTHA

35. Probability in Philosophy of Language 772
ERIC SWANSON

36. Decision Theory 789
LARA BUCHAK

37. Probabilistic Causation 815
CHRISTOPHER HITCHCOCK

Name Index 833
Subject Index 851

List of Contributors

John Aldrich, Economics Division, School of Social Sciences, University of Southampton

Guido Bacciagaluppi, Descartes Centre for the History and Philosophy of Science and the Humanities, Utrecht University; UMR 8590 IHPST, CNRS, Paris 1 and ENS; and UMR 7219 SPHERE, CNRS, Paris 7 and Paris 1

Paul Bartha, Department of Philosophy, University of British Columbia

David R. Bellhouse, Department of Statistical and Actuarial Sciences, University of Western Ontario

Lara Buchak, Department of Philosophy, University of California, Berkeley

Fabio G. Cozman, Engineering School, University of Sao Paolo

Vincenzo Crupi, Department of Philosophy and Education, University of Turin

Franz Dietrich, Paris School of Economics/CNRS and University of East Anglia

Antony Eagle, Department of Philosophy, University of Adelaide

Kenny Easwaran, Department of Philosophy, Texas A&M University

Terrence Fine, School of Electrical and Computer Engineering and Department of Statistical Sciences, Cornell University

Hans Fischer, Department of Mathematics, Catholic University of Eichstätt-Ingolstadt

James Franklin, School of Mathematics and Statistics, University of New South Wales

Roman Frigg, Department of Philosophy, Logic, and Scientific Method, London School of Economics

Maria Carla Galavotti, Department of Philosophy and Communication, University of Bologna

Donald Gillies, Department of Philosophy, University College London

Alan Hájek, School of Philosophy, Australian National University

James Hawthorne, Department of Philosophy, University of Oklahoma

Chris Hitchcock, Division of the Humanities and Social Sciences, California Institute of Technology

Stephen C. Hora, Management Science and Statistics, University of Hawaii at Hilo

Xia Jiang, Department of Biomedical Informatics, University of Pittsburgh School of Medicine

Matthew Kotzen, Department of Philosophy, University of North Carolina at Chapel Hill

Adam La Caze, Department of Philosophy, University of Queensland

Hannes Leitgeb, Munich Center for Mathematical Philosophy, Ludwig Maximilians University Munich

Christian List, Departments of Government and Philosophy, London, School of Economics

Aidan Lyon, Department of Philosophy, University of Maryland, College Park, and Munich Center for Mathematical Philosophy, Ludwig Maximilians University Munich

David McCarthy, Department of Philosophy, Hong Kong University

Roberta L. Millstein, Department of Philosophy, University of California, Davis

Wayne C. Myrvold, Department of Philosophy, University of Western Ontario

Richard Neapolitan, Division of Biomedical Informatics, Department of Preventive Medicine, Northwestern Feinberg School of Medicine

Wolfgang Schwarz, Department of Philosophy, University of Edinburgh

Michael Smithson, Research School of Psychology, Australian National University

Jan Sprenger, Tilburg Center for Logic and Philosophy of Science, Tilburg University

Eric Swanson, Department of Philosophy, University of Michigan

Edith Sylla, Department of History, North Carolina State University

Katya Tentori, Center for Mind/Brain Sciences; Department of Cognitive Sciences and Education, University of Trento

Michael Titelbaum, Department of Philosophy, University of Wisconsin, Madison

J. Robert G. Williams, Department of Philosophy, University of Leeds

Sandy Zabell, Department of Statistics, Northwestern University

Lyle Zynda, Department of Philosophy, Indiana University, South Bend

THE OXFORD HANDBOOK OF

PROBABILITY AND PHILOSOPHY

INTRODUCTION

ALAN HÁJEK AND CHRISTOPHER HITCHCOCK

Probability theory has long played a central role in statistics, the sciences, and the social sciences, and it is an important branch of mathematics in its own right. It has also been playing an increasingly significant role in philosophy—in epistemology, philosophy of science, ethics, social philosophy, philosophy of religion, and elsewhere. A case can be made that probability is as vital a part of the philosopher's toolkit as logic. Moreover, there is a fruitful two-way street between probability theory and philosophy: the theory informs much of the work of philosophers, and philosophical inquiry, in turn, has shed considerable light on the theory.

This volume encapsulates and furthers the influence of philosophy on probability, and of probability on philosophy. Nearly forty chapters summarize the state of play and present new insights in various areas of research at the intersection of these two fields. The chapters should be of special interest to practitioners of probability who seek a greater understanding of its mathematical and conceptual foundations, and to philosophers of probability who want to get up to speed on the cutting edge of research in this area. There is also plenty here to entice philosophical readers who don't work especially *on* probability but who want to learn more *about* it and its applications. Indeed, this volume should appeal to the intellectually curious generally; after all, there is much here to be curious about.

We do not expect all of this volume's audience to have a thorough training in probability theory. And while probability is relevant to the work of many philosophers, they often do not have much of a background in its formalism. With this in mind, we begin with "Probability for Everyone—Even Philosophers", a primer on those parts of probability theory that we believe are most important for philosophers to know. The rest of the volume is divided into seven main sections:

- *History*
- *Formalism*
- *Alternatives to Standard Probability Theory*
- *Interpretations*
- *Probabilistic Judgment and Its Applications*
- *Applications of Probability: Science*
- *Applications of Probability: Philosophy*

Some historians of probability, notably Hacking,[1] regard probability as having arrived surprisingly late on the intellectual scene, given its relative simplicity and practical value. Specifically, the birth of probability is usually dated to 1654, when Blaise Pascal and Pierre de Fermat began to correspond about a problem inspired by gambling on dice. (By contrast, Descartes' groundbreaking work in analytic geometry—a much more complex and abstract topic—appeared in 1637.) James Franklin's chapter in this volume traces the origins of probabilistic thinking much further back, even to antiquity. To be sure, the mid-to-late 17th century represents something of a watershed in the study of probability, and Edith Sylla takes up its history at that point, focusing on the work of Jacob Bernoulli and his influence in Continental Europe. Meanwhile, in Britain in the 17th and 18th centuries, probability theory was appropriated in various applications, as David Bellhouse details in his chapter. Hans Fischer continues the theory's history from early in the 19th century until around the middle of the 20th century, as probability increasingly became an autonomous branch of pure mathematics. During this period, statistics became an important field in England especially; this is the topic of John Aldrich's contribution to this volume. Finally, Maria Carla Galavotti canvases the work and legacy of some of the leading philosophers of probability in the 20th century, which created the field of philosophy of probability in its own right, and which set the stage for research in that field right up to today.

Two of the chief areas of research in the foundations of probability concern its *formalism*, and its *interpretation*. We begin with some suitable *formal theory* of probability, some codification of how probabilities are to be represented and how they behave. We then *interpret* that theory, bringing the formalism to life with an account of what probabilities are and of what grounds them. Regarding the formalism, Kolmogorov's axiomatization of 1933 remains orthodoxy. However, it has also found its share of critics. The chapters by Aidan Lyon and Kenny Easwaran discuss Kolmogorov's formalism and some of the sources of discontent with his approach. Richard Neapolitan and Xia Jiang's chapter on causal Bayes nets describes a newer formalism for efficiently representing and computing probabilities. The chapters by Terrence Fine, James Hawthorne, and Fabio Cozman describe formalisms intended as alternatives to standard probability theory, such as imprecise probabilities and qualitative analogues of probability. J. Robert G. Williams discusses how Kolmogorov's approach may be generalized to accommodate various nonclassical logics.

Regarding the interpretation of probability, we are pulled in multiple directions. Probability apparently begins in uncertainty, but it arguably does not end there. We are irremediably ignorant of various aspects of the world; probability theory has been our chief tool for systematizing and managing this ignorance. Our evidence is impoverished, and it typically fails to settle various matters of interest to us. But even if Hume was right that there are no necessary connections between distinct existences, still it seems that there are *probabilistic* connections between the evidence that we have and the hypotheses that we entertain. Moreover, many authors believe that modern physics gives us reason to think that the world itself has not settled various matters either: probability is part of the fabric of reality. If this is right, then *pace* Einstein, God *does* play dice.

Accordingly, philosophers have homed in on three leading kinds of probability: evidential, subjective, and physical. (Note that one can consistently adhere to more than

[1] *The Emergence of Probability, A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. 1975. Cambridge: Cambridge University Press.

one interpretation of probability. Indeed, a case can be made for embracing evidential, subjective, and physical probabilities for different purposes.) *Evidential probability* is meant to capture the degree to which available or hypothetical evidence supports various hypotheses, where typically such support falls short of entailment. One might think of this as how objectively plausible the hypotheses are in light of the evidence, irrespective of what anyone actually thinks. The earliest incarnation of this idea was enshrined in the *classical* interpretation of probability, in which the probability of an event is regarded as the ratio of the number of live possibilities favourable to the event divided by the total number of live possibilities. (The *live* possibilities are those that have not been ruled out.) This interpretation is founded on the *indifference* principle: when there is no relevant evidence, or when the relevant evidence bears symmetrically on the alternative possibilities, the possibilities should be given equal weight. Sandy Zabell's chapter on symmetry arguments in probability traces the history of the indifference principle, and more recent heirs of the classical interpretation that aim to ground probability values in symmetries.

The *logical* interpretation of probability generalizes this approach to evidential probability, seeking to measure the degree of support that a body of evidence gives a particular hypothesis, whatever the evidence and hypothesis. The possibilities can be assigned *unequal* weights, and probabilities can be computed whatever the evidence may be, symmetrically balanced with respect to the hypothesis or not. The result, to the extent that it succeeds, is a comprehensive *inductive logic* or *confirmation* theory.[2] Several chapters in this volume discuss at least to some extent these themes of evidential probability, the classical and logical interpretations of probability, and confirmation theory. Maria Carla Galavotti's chapter discusses the logical interpretation of probability in the early 20th century, with particular attention to the work of Harold Jeffreys. Vincenzo Crupi and Katya Tentori focus on confirmation theory and inductive logic in their chapter. Matthew Kotzen also addresses evidential probability in his contribution, paying special attention to the work of Kyburg and Williamson. Sandy Zabell's discussion of symmetry arguments additionally covers many ideas that have been associated with the logical interpretation of probability.

Some authors in philosophy of probability's pantheon, however, were sceptical of any notion of logical probability—notably, Ramsey and de Finetti. They advocated a more permissive *subjectivism* about probability, which interprets probabilities as degrees of confidence of suitably rational agents. This interpretation is addressed especially by Lyle Zynda's chapter, but it also takes centre stage in the chapters by Fabio Cozman, Franz Dietrich and Christian List, Stephen Hora, Michael Smithson, and Jan Sprenger.

Meanwhile, a number of authors hold that probabilities reside in the world itself, mind-independently—these are *physical* probabilities, often called *chances*.[3] Frequentist interpretations identify such probabilities with appropriate relative frequencies in some sequence of events; see the chapter by Adam La Caze. These interpretations seem to fare better when there are many trials of the relevant event type—flips of a coin, throws of a die,

[2] Terminological caution: in ordinary English, "confirms" usually means *establishes* or *verifies*, but confirmation theory's relations are *probabilistic*. Moreover, these relations include those of evidential *counter*-support.

[3] They are sometimes also called *objective* probabilities. However, logical probabilities are often regarded as objective also (much as logic itself is often regarded as objective). We thus prefer to speak of *physical* probabilities.

and the like. Frequentist interpretations fare worse when there are few trials, and especially poorly when there is just one—this is 'the problem of the single case'. Partly as a response to this problem, propensity interpretations regard probabilities as graded dispositions or tendencies—applicable even in the single case (at least on some versions). Donald Gillies surveys these interpretations. More recently, *best systems* approaches to physical probability have become popular. On this type of view, chances fall out of the theory of the universe that best balances certain theoretical virtues. This is the topic of Wolfgang Schwarz's chapter.

There are further interpretive issues regarding probabilities that are not specifically interpretations *of* probability, although they interact with these interpretations. Probability has been thought by many authors to bear interesting connections to *randomness*, the subject of Antony Eagle's chapter. There is also considerable debate about whether chance is compatible with *determinism*, an issue that Roman Frigg takes up. Issues connected to the interpretation of probability also underlie the debate between champions of Bayesian and frequentist approaches to statistical inference —see Jan Sprenger's contribution.

We then turn to probabilistic judgment and its applications. Michael Smithson explores some of the psychological literature on how people reason with probabilities. While it is often useful to represent the opinion of someone, especially an expert, in probabilistic form, many of us cannot simply assign a number to the degree of our conviction on the basis of introspection. Stephen Hora describes a number of strategies for eliciting probabilities from subjects. Franz Dietrich and Christian List consider how the probabilistic judgments of a number of individuals can be pooled in various ways.

Next, our authors discuss applications of probability, beginning with science. Physics explicitly traffics in probabilities, especially in quantum mechanics and statistical mechanics. Guido Bacciagaluppi and Wayne Myrvold, respectively, examine the place of probabilities in each of these theories. Furthermore, probabilities make their way both explicitly and implicitly into biology, especially in connection with the concept of fitness in evolutionary biology—see Roberta Millstein's chapter.

While the earlier sections on formalism and interpretations fall under the *philosophy of probability*, the final section concerns the myriad applications of *probability in philosophy*. Many areas of philosophy have benefited from probability theory. Several chapters display this: Matthew Kotzen's on epistemology; Hannes Leitgeb's on logic; David McCarthy's on ethics; Paul Bartha's on philosophy of religion; and Eric Swanson's on philosophy of language. A number of chapters survey more targeted applications of probability in philosophy: Vincenzo Crupi and Katya Tentori's on confirmation theory; Michael Titelbaum's on self-locating credences; Lara Buchak's on decision theory; and Christopher Hitchcock's on probabilistic causation.

Stepping back, we see just how fertile the interaction of probability and philosophy can be. This is an exciting time for the philosophy of probability, and probability theory's value to philosophy has never been as appreciated as it is nowadays. We thus thought it was especially timely when Peter Momtchiloff of Oxford University Press approached us with the idea of this Handbook. He has been a pleasure to work with, and we thank him for his ongoing encouragement and advice. Many thanks are also due to John Cusbert and Edward Elliott for their incisive comments on various drafts from the authors. Above all, we thank the authors themselves for their fine work.

CHAPTER 1

PROBABILITY FOR EVERYONE—EVEN PHILOSOPHERS

ALAN HÁJEK AND CHRISTOPHER HITCHCOCK

1.1 Introduction

Many people who would benefit from knowing some probability theory have little or no background in it. This is especially true of philosophers interested in areas to which it is relevant, but intimidated by discussions that turn on it. They have nothing to fear: the basics of probability theory are remarkably simple. Most philosophical applications of probability require no more than elementary set theory, arithmetic, and high-school algebra. Its entire axiomatization can be understood with just a little calculus. To be sure, more advanced probability theory appeals to measure theory, Lebesgue's theory of integration, advanced set theory, non-standard analysis, and more. However, philosophical debates enter such arcane territory relatively rarely. The required mathematics is mostly painless!

This chapter surveys the parts of probability theory that we believe are most useful to know—especially for philosophers. We begin with its axiomatization, in two passes: first, a more informal treatment of the elementary theory that is suitable for finite applications, and some of its most important theorems; second, a more formal treatment of the more sophisticated theory that is suitable for infinite applications, and some further theorems. (Readers who are familiar with the basics but who want a primer or a refresher of the more sophisticated theory may want to skip the first pass.) We also introduce some further key probabilistic concepts, especially those that are philosophically significant. Of special interest to philosophers are several probabilistic paradoxes that we present—the first pass already suffices for them.

1.2 A First Pass: Elementary Probability Theory

Probabilities are numerical values that are assigned to 'events', or 'propositions', by a probability function P—we will speak both ways. These bearers of probabilities are usually understood as sets of possible outcomes of some random experiment, or sets of worlds, though probability theory itself does not take a stand on this. Either way, they are sets of possibilities belonging to some 'universal set' $\mathbf{\Omega}$, the set of all possible outcomes, or worlds. Let A and B be events, so understood. Probabilities obey the following axioms. They are non-negative:

$$P(A) \geq 0.$$

The probability of the universal set is 1:

$$P(\Omega) = 1.$$

And they are additive: the probability that one of two mutually exclusive events occurs is the sum of their individual probabilities:

$$P(A \text{ or } B) = P(A) + P(B) \text{ if } A \text{ and } B \text{ cannot both occur.}$$

We will call this property *finite additivity*, to distinguish it from a stronger form of additivity to be discussed later. For example, for the random experiment of tossing a fair die once and observing how it lands, a natural universal set would be {1, 2, 3, 4, 5, 6}. Assume that each outcome has probability 1/6. The three different ways that the die can land with an even number showing up (2, 4, and 6) are mutually exclusive, so

$$P(\text{die lands even}) = 1/6 + 1/6 + 1/6 = 3/6 = 1/2.$$

A theorem that follows immediately from our axioms is the *negation rule*:

$$P(\text{not-}A) = 1 - P(A).$$

Hence, for our die experiment, P(die lands odd) = 1 – P(die lands even) = ½.

When A and B are not mutually exclusive, we may still determine the probability of the disjunction A or B, using the additivity axiom:

$$P(A \text{ or } B) = P(A \text{ and not-}B) + P(\text{not-}A \text{ and } B) + P(A \text{ and } B). \quad (*)$$

Note also that

$$P(A) = P(A \text{ and } B) + P(A \text{ and not-}B)$$

and

$$P(B) = P(A \text{ and } B) + P(\text{not-}A \text{ and } B).$$

Substituting into (∗) and rearranging, we achieve this important generalization of the additivity axiom:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B).$$

Note that these axioms and theorems also hold of various non-probabilistic quantities—e.g. length, area, volume, and mass, normalized (scaled by a factor) so that there is a maximal value of 1. This gives us various ways to represent probabilities pictorially. For example, following van Fraassen (1989), we may think of A, B, and so on as regions on a Venn diagram, and think of probabilities as amounts of mud heaped on the regions; the total amount of mud over the entire diagram is 1 unit. The amount of mud on any region is non-negative, and the total amount of mud covering two (or more) non-overlapping regions is the sum of their individual amounts.

1.3 Conditional Probability

Conditional probability is probability relativized to, or restricted to, some particular event or proposition (usually regarded as a potential piece of information or evidence). The *conditional probability of A given B* is given by the ratio of unconditional probabilities:

$$P(A|B) = P(A \text{ and } B)/P(B), \text{ provided } P(B) > 0.$$

Thus, the probability that our fair die lands 2 is 1/6; but the conditional probability that it lands 2, *given* that it lands even, is 1/3:

$$P(\text{die lands 2}) = 1/6;\ P(\text{die lands 2}|\text{die lands even}) = \frac{1/6}{3/6} = 1/3.$$

We now have easy ways to compute the probability of a conjunction when the requisite conditional probabilities are available:

$$P(A \text{ and } B) = P(A|B)\,P(B) = P(B|A)\,P(A).$$

A useful theorem is the *law of total probability*, which expresses an unconditional probability as a weighted average of conditional probabilities:

$$P(A) = P(A|B)\,P(B) + P(A|\text{not-}B)\,P(\text{not-}B).$$

We may generalize this result to an arbitrary finite partition $\{B_1,\ldots, B_n\}$—a *partition* is a set of mutually exclusive events that collectively exhaust the space of possibilities. Put another way, if $\{B_1,\ldots, B_n\}$ is a partition, then every possibility in Ω belongs to exactly one B_i. In this case:

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

when all of the conditional probabilities are defined.[1]

[1] If some elements of the partition have probability 0, we can sum over all elements that have positive probability.

1.4 Bayes' Theorem

A particularly famous result involving conditional probabilities is *Bayes' theorem*. Here's an elegant formulation:

$$\frac{P(A|B)}{P(A)} = \frac{P(B|A)}{P(B)}$$

assuming here, and in what follows, that the conditional probabilities are defined. More commonly used is the following formulation; our notation evokes a standard usage of Bayes' theorem in determining the probability of a *hypothesis H* in light of *evidence E*:

$$\begin{aligned} P(H|E) &= \frac{P(E|H)\,P(H)}{P(E)} \\ &= \frac{P(E|H)\,P(H)}{P(E|H)\,P(H) + P(E|\text{not-}H)\,P(\text{not-}H)} \end{aligned}$$

More generally, if $\{H_1,\ldots, H_n\}$ is a partition,

$$P\left(H_i|E\right) = \frac{P\left(E|H_i\right)P\left(H_i\right)}{\sum_{j=1}^{n} P\left(E|H_j\right)P\left(H_j\right)}$$

Bayes' theorem allows us to see how the so-called *posterior probability*, $P(H|E)$, depends on three things:

- the *likelihood,* $P(E|H)$;
- the *prior* probability: the probability of the hypothesis, $P(H)$ (antecedent to any relevant evidence); and
- the probability of the evidence, $P(E)$.

Some cautions about the terminology: In ordinary English, the word 'likelihood' sounds synonymous with 'probability', but in this technical usage it is a very specific conditional probability. Talk of 'prior' and 'posterior' probability should also be handled with care (which unfortunately is often not the case in the literature). It suggests something diachronic: a temporal passage from an earlier probability function to a later one. But Bayes' theorem is entirely synchronic: it involves different aspects of a single probability function *P*, and as such it is on a par with all the other (synchronic) theorems of probability theory. The function *P* that appears on the left-hand side of the equation is the *same* as the one that appears multiple times on the right-hand side—it is just as 'prior' (or otherwise) wherever it appears!

This way of talking stems from an intended interpretation of *P* as the probability function that represents some agent's degrees of confidence—*credences*—regarded as assignments of probabilities to various propositions. Let the agent be *you* in an initial—yes, 'prior'—state of mind. Among other things, your degree of confidence in *H* is $P(H)$. Now suppose that you learn a piece of evidence *E*, and that this is the strongest piece of information of which you become certain. What should your *new* degrees of confidence look like—that is, how

should you *update P*? The favoured rule among so-called 'Bayesians' is *conditionalization*:

$$\text{For all } X, P_{\text{new}}(X) = P(X|E) \text{ (Conditionalization)}$$

In particular, your new probability of H is $P(H|E)$. *This* involves a diachronic process, a change in your probability function from one time to a later time. For example, suppose that a die is about to be tossed, and you initially assign probability 1/6 to each of the 6 possible outcomes. Your probability, then, that the die lands 2 equals 1/6. Now suppose that the die is tossed; you don't learn the exact outcome, but you do learn that the die landed even (which you initially gave probability ½). Then your new probability for the die landing 2 should be:

$$P(\text{die lands 2}|\text{die lands even}) = \frac{1/6}{1/2} = 1/3.$$

Conditionalizing on E takes us from one probability function to a new one that resembles it as much as possible, while confining all of its probability to E, and renormalizing so that the total probability assigned to E is 1. We may picture this with a 'muddy Venn diagram' in which all the mud assigned to not-E is scraped away, and the mud that remains is then regarded as 1 unit in total.

The rule of conditionalization is sometimes erroneously called 'updating by Bayes' theorem'. Nothing in the rule requires any appeal to that theorem. To be sure, sometimes the theorem might be helpful in calculating the requisite conditional probability on the right-hand side of (Conditionalization). Then again, it might not be—the simple ratio formula often suffices for the calculation, and the conditional probability might be available without any calculation at all. Notice that we made no appeal to Bayes' theorem in the example just given.

Bayes' theorem provides a salutary reminder of the importance of the (so-called) prior in the computation of a conditional probability. Consider a kind of question at which normal subjects are notoriously bad. A test for a disease is 99% effective, by which we mean:

$$P(\text{test yields positive result}|\text{subject has disease}) = 0.99,$$

and

$$P(\text{test yields negative result}|\text{subject does not have disease}) = 0.99.$$

Suppose a randomly selected person takes the test, and tests positive. What is the probability that this person has the disease? You may be tempted to say: 0.99. But we have not given enough information to answer this question; we need to know the 'base rate' of the disease, the prior probability that a randomly selected person has it. Suppose that this probability is 1%; then the answer is 1/2. Now suppose that the prior probability is 0.1%; then the probability conditional on a positive test result is about 0.09.[2] This may seem startling, given that the test looks highly reliable. Bayes' theorem reminds us that when the antecedent probability of having the disease is very small, the positive result is more likely to be a *false* positive than a true positive. This is most easily seen by imagining a population of 1,000 people who take the test, only one of whom has the disease. The test is very likely to produce

[2] The calculation for the last figure is: $\frac{0.99 \times 0.001}{(0.99 \times 0.001) + (0.01 \times 0.999)}$.

a positive result for that one person. But it is also likely to produce around 10 positive results among the 999 people who don't have the disease. Given these figures, only 1 in 11, or 0.09, of those who test positive have the disease. The failure to pay attention to the base rate in calculating a posterior probability is called the *base rate fallacy*.

1.5 Independence

If $P(A \text{ and } B) = P(A)P(B)$, then A and B are said to be *independent*. This factorization characterization may seem surprising at first, but there is a simple rationale behind it. Assuming that $P(B) > 0$, this is equivalent to

$$P(A|B) = P(A).$$

Similarly, if $P(A) > 0$, A and B are independent just in case

$$P(B|A) = P(B).^{3}$$

The latter two equations capture the intuitive idea: when two events are independent, the occurrence of one of the events is completely uninformative about the occurrence of the other, probabilistically speaking. For instance, successive tosses of a die, or a coin, or successive spins of a roulette wheel are typically regarded as independent.

Two cautions. Firstly, the locution 'A is independent of B' is somewhat careless, encouraging one to forget that independence is a relation that events or sentences bear *to a probability function*. Probabilistic independence is really a three-place relation—and as such, it is the odd one out among various notions of independence—causal, counterfactual, logical, metaphysical, etc. Secondly, for this and other reasons, probabilistic independence should not be identified with causal independence, counterfactual independence, or any other pre-theoretical sense of the word, even though such identifications are often made in practice. For example, what basis do we have for assuming, as we typically do, that heads on successive tosses of a coin are *probabilistically* independent events? To be sure, we may safely assume that they are causally independent, counterfactually independent, and so on—but that is another matter. Moreover, in cases where an agent gains evidence about a coin that she thinks may be biased, successive tosses may be probabilistically *dependent* by her lights, while being causally independent, counterfactually independent, and so on. Suppose you give some credence to the coin being fair, but also some credence to its being two-headed; you toss it, and it lands heads. This is some evidence for you that the coin is two-headed, and it thus should increase your confidence that the next toss will land heads.

If $P(A \text{ and } B) \neq P(A)P(B)$, then A and B are said to be *dependent* or *correlated*.
If $P(A \text{ and } B) > P(A)P(B)$, then A and B are said to be *positively correlated*.
If $P(A \text{ and } B) < P(A)P(B)$, then A and B are said to be *negatively correlated*.

[3] Note that if $P(A) = 0$ or $P(B) = 0$, these equivalences break down—see Fitelson and Hájek (2014).

If A and B are positively (negatively) correlated, then we say that each event is *evidence for* (*against*) the other, and that each event *confirms* (*disconfirms*) the other. For example, suppose that A and B are positively correlated. If you were to learn B (and nothing else), and update your probabilities by conditionalization (as described in the previous section), the probability of A would increase. If you were to learn A (and nothing else), your probability for B would increase. We may also define a three-place evidential relation:

> *E favours H over G* just in case $P(E|H) > P(E|G)$.

That is, E is rendered more likely by H than it is by G.

We can also extend the notions of independence and correlation to probabilities *conditional on* some event C. If $P(A \text{ and } B|C) = P(A|C)P(B|C)$, then A and B are said to be *independent, conditional on C*. If $P(A \text{ and } B|E) > P(A|E)P(B|E)$, then A and B are said to be *positively correlated conditional on E*; reverse the inequality for their being *negatively correlated conditional on E*.

This completes our first pass. It is remarkable how powerful this elementary probability theory is. It suffices for most philosophical applications. Before turning to a more sophisticated development of probability in terms of measure theory, we will examine some simple puzzles and paradoxes in probability that require no more than elementary probability theory.

1.6 Probability Puzzles and Paradoxes

These are not true paradoxes, in the philosophical sense—puzzles without clear answers, in which no answer seems satisfactory, or in which incompatible answers each seem correct. Rather, they are problems in which the mathematics is well understood, but probabilities behave in unintuitive ways.[4]

1.6.1 Simpson's Paradox

Simpson's paradox is named after Edward Simpson, who gave a particularly clear and compelling exposition of it in 1951 (Simpson 1951). But it was known well before this, and described in Yule (1903).

Simpson's paradox arises when two events are (positively or negatively) correlated conditional on every member of a partition, but they are unconditionally independent or have the opposite correlation. Formally, let A and B be events; we have a case of Simpson's paradox if there exists a partition $\{C_1, C_2, \ldots, C_n\}$ such that:

i) $P(A \text{ and } B|C_i) > P(A|C_i)P(B|C_i)$ for each i, but $P(A \text{ and } B) \leq P(A)P(B)$;
ii) $P(A \text{ and } B|C_i) < P(A|C_i)P(B|C_i)$ for each i, but $P(A \text{ and } B) \geq P(A)P(B)$;

[4] We might consider the oft-neglected role of base rates in calculating posterior probabilities, which we have already seen, as 'paradoxical' in this sense.

iii) $P(A \text{ and } B|C_i) \geq P(A|C_i)P(B|C_i)$ for each i, but $P(A \text{ and } B) < P(A)P(B)$; or
iv) $P(A \text{ and } B|C_i) \leq P(A|C_i)P(B|C_i)$ for each i, but $P(A \text{ and } B) > P(A)P(B)$.

These cases are not mutually exclusive; for example, a case in which $P(A \text{ and } B|C_i) > P(A|C_i)P(B|C_i)$ for each i, but $P(A \text{ and } B) < P(A)P(B)$, would be of types (i) and (iii).

Simpson's own example was of type (i), where equality holds unconditionally ($P(A \text{ and } B) = P(A)P(B)$). A strict use of the name 'Simpson's paradox' would refer to this type of case only, but it is common to use the name to refer to all cases of types (i) – (iv).

Here is a hypothetical example. Suppose we conduct a study on whether eating chocolate (C) is correlated with hypertension (H). Everyone in the study is either female or male (F or M, with a subject being F if and only if they are not-M).[5] We find that among women, higher rates of hypertension prevail among those who eat chocolate: $P(H|F \text{ and } C) = 0.3 > 0 = P(H|F \text{ and not-}C)$. Similarly, among men, those who eat chocolate have higher rates of hypertension: $P(H|M \text{ and } C) = 0.8 > 0.5 = P(H|M \text{ and not-}C)$. Nonetheless, in the total population, chocolate eaters are no more likely to suffer from hypertension than those who abstain: $P(H|C) = P(H|\text{not-}C) = 0.4$.

How is this possible? It can happen if women are more likely to eat chocolate. Overall (in this hypothetical study), women have lower rates of hypertension than men, and women are more likely to eat chocolate than men. This cancels out the correlation between hypertension and eating chocolate that exists in each subgroup. For example, if the study contains 100 subjects, the total numbers could look like this:

	Chocolate hypertension / total	No Chocolate hypertension / total
Men	8 / 10	20 / 40
Women	12 / 40	0 / 10
Total	20 / 50	20 / 50

This is a case of type (i), where equality holds unconditionally. It is also possible for the overall rate of hypertension to be lower for chocolate eaters in the combined population (just increase the difference in hypertension rates between men and women, or increase the percentage of women who eat chocolate).

There are many real-life examples of Simpson's Paradox. For example, Westbrooke (1998) analysed data collected by Statistics New Zealand on the representation of ethnic Maoris in jury pools. Overall, the representation of Maoris in jury pools was slightly higher than their representation in the population as a whole (0.1 vs. 0.095). But when the statistics were broken down by district, it turned out that Maoris were under-represented in jury pools in every single district. The over-representation in New Zealand as a whole occurred because districts with large Maori populations tended to have a higher proportion of the population participating in jury pools. This is a case of types (ii) and (iv), where A represents being Maori, B represents service in a jury pool, and the C_i's represent one's district of residence.

[5] We recognize that neither biological sex nor gender identity divide cleanly into the dichotomy of male and female. We apologize to those who were excluded from this hypothetical study. The good news is that they can eat all the chocolate they like.

While the mathematics behind Simpson's Paradox is well-understood, it can seem paradoxical. Pearl (2009) argues that the appearance of paradox arises because we often want to interpret correlations causally. Consider our original example involving chocolate and hypertension. Suppose that we are considering whether to issue a warning against the consumption of chocolate. If we interpret the correlation (or lack thereof) between chocolate consumption and hypertension causally, we will conclude that chocolate causes hypertension in both men and women, but that it does not cause hypertension in the population of men and women as a whole. This is clearly impossible. This leaves us with the question of how to partition the data in order to draw causal conclusions. If we want to know whether chocolate causes hypertension, should we look at the correlation among men and women separately, or should we look at the lack of correlation in the population as a whole? In this case, the answer is that we should look at the data broken down by sex. But it is not always right to break down the data into sub-populations. The details depend upon the causal structure of the case.

Note, however, that Simpson-style cases may seem paradoxical to us even *without* our interpreting correlations causally: it is already unintuitive that *proportions* can behave in these ways. Replace the chocolate example with one in which there is no temptation to adopt a causal interpretation. Consider an urn with 100 balls that are made either of copper or nickel, that are either maroon or white, and that are either hard or soft. Replace '**m**en' by '**m**aroon', '**w**omen' by '**w**hite'; '**c**hocolate' by '**c**opper', and '**n**o chocolate' by '**n**ickel'; '**h**ypertension' by '**h**ard' in the table above. It is still surprising that: (i) among the maroon balls, a higher proportion of the copper balls than of the nickel balls are hard (0.8 vs. 0.5); (ii) among the white balls, a higher proportion of the copper balls than of the nickel balls are hard (0.3 vs. 0.0); yet (iii) the overall proportion of hard copper balls is the same as the proportion of hard nickel balls. We are not understanding these correlations causally; we just have bad arithmetical intuitions. The 'paradox' in probability is really based on a 'paradox' in arithmetic.

1.6.2 Berkson's Paradox

This paradox was formulated by Joseph Berkson (1946). It is similar to Simpson's Paradox in that it involves differences in the behaviour of conditional and unconditional probabilities. Two events A and B can be unconditionally independent, but nonetheless be (positively or negatively) correlated conditional on a third event, C. Formally:

$$P(A \text{ and } B) = P(A)\,P(B), \text{but } P(A \text{ and } B|C) \neq P(A|C)\,P(B|C).$$

Berkson was specifically concerned with the case where C is a condition for inclusion in some study or data set, and A and B influence whether C occurs. For instance, suppose that A and B are independent diseases, each sufficiently serious to warrant admission to a hospital (event C). If you conduct a study of hospital patients to determine whether A and B are correlated, you may find a negative correlation, even though A and B are independent in the population as a whole. Among those who are in the hospital, those who do not have disease A all have some other reason for being there, and thus are more likely to have disease B.

Here is another hypothetical example: Suppose that in the population as a whole, 20% are intelligent, 20% are attractive, and these two traits are independent. These two traits influence whether one appears on television. Most people have only a 5% chance of appearing on television. Intelligent but unattractive people have a 20% chance of appearing on television; attractive but unintelligent people have a 50% chance; and those who are doubly blessed have a 60% chance. In a population of 1,000 people, the numbers would look like this:

	Attractive	Unattractive	Combined
	On TV / total	On TV / total	On TV / total
Intelligent	24 / 40	32 / 160	56 / 200
Unintelligent	80 / 160	32 / 640	112 / 800
Combined	104 / 200	64 / 800	168 / 1000

Among the 168 people who appear on television, 56 or 33% are intelligent; 104 or 62% are attractive; and 24 or 14% are both attractive and intelligent. Since $0.14 < 0.33 \times 0.62 \approx 0.21$, intelligence and attractiveness are negatively correlated among those who appear on television. It is not surprising that if you stay at home and watch television, never going out in the world to meet real people, you will overestimate the proportion of attractive people in the population. It is somewhat more surprising that you will also conclude that attractive people tend to be less intelligent, even though this is not really the case. Interestingly, intelligence and attractiveness are also negatively correlated among those who are *not* on television, although in this case the numbers are much closer (.019 vs. .02). This case is thus an instance of Simpson's Paradox of type (ii) described above.

Berkson's paradox is a special case of *sampling bias* or *selection effect*, where the proportions that occur within a sample are not representative of a population as a whole. Such sampling bias in general is familiar: for example, if you wanted to estimate the average height of students in a school based on a sample of students, you should not use the members of the boys' and girls' basketball teams as your sample; you have every reason to believe that they will be taller than the average students. Berkson's paradox is a subtler example, where the sampling method affects the correlation that is found between two different traits.

1.6.3 The Monty Hall Paradox, and Related Puzzles

The Monty Hall Paradox is named after the creator and former host of the game show 'Let's Make a Deal'. The problem has led to a great deal of heated discussion, as a Google search shows. Imagine that you are a contestant on the show. You are shown three doors, numbered 1, 2, and 3. Behind one of the doors is a valuable prize—say, a sports car. Behind the other two doors are 'zonks' or booby prizes (perhaps a baby hippopotamus and a wheelbarrow full of jello). You can choose one door, and will receive whatever is behind it. You believe that each door is equally likely to hide the prize, so you randomly pick door 1 (say). Monty Hall, the host, then builds suspense by showing you what is behind one of the doors you didn't

choose. Suppose he reveals a baby hippo behind door 2. You believe that Monty Hall knows which door hides the real prize, and that he deliberately revealed a zonk (if he revealed the car, it would not build suspense). He then offers you a choice: you can stick with door 1, or switch to door 3. What should you do?

It may seem as if it doesn't matter. Each door had an equal probability of hiding the prize, and you knew all along that at least one of the doors you didn't choose would hide a zonk. But it turns out that you should switch: door 3 now has a 2/3 probability of winning, while door 1 has only a 1/3 probability of hiding the prize.

We can show this using Bayes' theorem. Assume that you have initially chosen door 1. Let C_1 be the event corresponding to the car being behind door 1; C_2 and C_3 are defined analogously. Let R_2 be the event corresponding to Monty Hall revealing a zonk behind door 2. Obviously, $P(C_2|R_2) = 0$. But we must calculate $P(C_1|R_2)$ and $P(C_3|R_2)$. From the structure of the game, we know that the prior probabilities are

$$P(C_1) = P(C_2) = P(C_3) = 1/3.$$

We also know that

$$P(R_2|C_1) = 1/2, P(R_2|C_2) = 0, \text{and } P(R_2|C_3) = 1.$$

If the prize is behind door 1, then both doors 2 and 3 hide zonks. Monty Hall could show either of these doors to build suspense, and we don't have any reason to believe that he would show one door rather than the other. It's important here to note that R_2 represents Monty Hall's *showing* you a zonk behind door 2. R_2 does not represent 'there *is* a zonk behind door 2' (which is represented by not-C_2). If R_2 had the latter meaning, we would have $P(R_2|C_1) = 1$ instead of 1/2. If the car is behind door 2, then Monty Hall definitely will not open door 2 (he will show us the zonk behind door 3 instead). If the car is behind door 3, then he will be forced to show us the zonk behind door 2 as a way of building suspense. Plugging the corresponding probabilities into Bayes' theorem, we get:

$$\begin{aligned}
P(C_1|R_2) &= [P(C_1)P(R_2|C_1)]/P(R_2) \\
&= \frac{P(C_1) \times P(R_2|C_1)}{[P(R_2|C_1) \times P(C_1)] + [P(R_2|C_2) \times P(C_2)] + [P(R_2|C_3) \times P(C_3)]} \\
&= \left[1/3 \times 1/2\right] / \left(\left[1/2 \times 1/3\right] + \left[0 \times 1/3\right] + \left[1 \times 1/3\right]\right) \\
&= \left[1/6\right] / \left[1/6 + 1/3\right] \\
&= 1/3
\end{aligned}$$

And

$$\begin{aligned}
P(C_3|R_2) &= 1 - P(C_1|R_2) \text{ (since } P(C_2|R_2) = 0) \\
&= 2/3.
\end{aligned}$$

The key to solving this problem is realizing that Monty Hall is more likely to show you what's behind door 2 if the prize is behind door 3 than if it is behind door 1. If it is behind door 3, he has to open door 2; but if it is behind door 1, he could open either door 2 or door 3. In other words, the *likelihoods* are different: $P(R_2|C_1) = 1/2$ and $P(R_2|C_3) = 1$. This means that when he reveals the zonk behind door 2, he gives you evidence that confirms the

hypothesis that the prize is behind door 3. In particular, the information that you get when he shows you the zonk behind door 2 is not merely that door 2 does not contain the prize (not-$C2$). A quick computation shows that $P(C1|\text{not-}C2) = P(C3|\text{not-}C2) = 1/2$. Also, if Monty Hall does not know which door hides the prize, and picked door 2 at random, then it would no longer be advantageous to switch doors. Arriving at the correct answer thus requires that one understand his procedure to choose which door to open.

Here's the simplest way to see that switching is advantageous. Suppose that you commit to a strategy of 'sticking' or 'switching' before you choose a door. It is easy to see that if you pursue a strategy of sticking, you will win the prize just in case your first guess is correct. For instance, if you initially choose door 1, and then stick with door 1 when offered a chance to switch, then you will win just in case the prize is behind door 1. Thus a strategy of sticking has a one in three probability of success. By contrast, switching will win you the prize just in case your initial choice was *incorrect*. For example, suppose you choose door 1, and the prize is behind door 2. Monty Hall will now show you what's behind door 3, and switching to door 2 will win you the prize. Parallel reasoning applies if the prize is behind door 3. So a strategy of switching will win the prize in two cases out of three. This is borne out by various simulations of the Monty Hall game that can be found on the internet.

As Glymour (2001) points out, the Monty Hall Paradox is actually an instance of Berkson's Paradox. To see this, we need to consider another event. Let $G1$ represent that you initially guessed that the prize is behind door 1, and likewise for $G2$ and $G3$. Let us assume that you guess randomly, so that $P(G1) = 1/3$. Since you have no idea where the prize is, $G1$ is probabilistically independent of $C1$, $C2$, and $C3$. However, your initial choice and the actual location of the prize both influence which door Monty Hall chooses to open. As a result, your initial guess and the location of the prize are negatively correlated conditional on his choice. That is:

$$P(G1 \text{ and } C1) = P(G1)\,P(C1), \text{ but } P(G1 \text{ and } C1|R2) < P(G1|R2)\,P(C1|R2).$$

There are a number of other puzzles and paradoxes that have essentially the same structure as the Monty Hall Paradox. For example, the Three Prisoners Paradox, described by Gardner (1959), is mathematically identical to the Monty Hall problem. A closely related puzzle was presented by Joseph Bertrand (1889). See Sandy Zabell's chapter, this volume (15), for discussion of the latter two puzzles.

1.7 A Second Pass: More Sophisticated Probability Theory

Let us now move on to our second pass, following Kolmogorov in giving probability theory a set-theoretic underpinning, and strengthening the axiomatization to make it fit for infinitary applications. This will be needed for some of the great limit theorems of probability theory, and for more advanced philosophical applications. The *locus classicus* of the mathematical theory of probability is Kolmogorov (1933/1950), who found his

inspiration in measure theory. His axiomatization has become orthodoxy.[6] Things get a little harder at this point, but still they can be presented in a user-friendly way.

1.7.1 Kolmogorov's axioms

Start with a non-empty set Ω. A *field* (*algebra*) on Ω is a set $\mathscr{F}$ of subsets of Ω that has Ω as a member, and that is closed under complementation (with respect to Ω) and union.[7]

The members of $\mathscr{F}$ are called the *measurable sets*—an homage to measure theory, under which probability theory will be subsumed. It would be more appropriate to call them the *measured* sets—those that *are assigned values* by the relevant probability function. We may think of them as the 'well-behaved' sets from the point of view of that function. Philosophers may think of them as the privileged sets that are fit to be the bearers of probability assignments.

The members of $\mathscr{F}$ are often called 'events', as before. It should be noted, however, that this is again a technical usage of the term, rather far removed from the commonsensical usage of that word, when we speak of *events* such as the French Revolution, or the assassination of Kennedy. *Events* in that sense do not have the closure properties of an algebra. For example, plausibly there is no such *event* as the *non-assassination of Kennedy*; that would be too heterogeneous, too gerrymandered to count, since it would be instantiated in so many disparate ways (Kennedy never existed; the bullets were never fired; the bullets were fired but Kennedy survived them…). Nor is there such an *event* as *the French Revolution OR the assassination of Kennedy.* Yet probability theory takes such combinations of its objects in its stride.

Assume for now that Ω is finite. Let P be a function from $\mathscr{F}$ to the real numbers obeying:

1. $P(A) \geq 0$ for all $A \in \mathscr{F}$.
2. $P(\Omega) = 1$.
3. $P(A \cup B) = P(A) + P(B)$ for all $A, B \in \mathscr{F}$ such that $A \cap B = \emptyset$.

Call P a *probability function*, and $(\Omega, \mathscr{F}, P)$ a *probability space*.

Kolmogorov extends his axiomatization to cover infinite probability spaces—ones for which Ω is infinite. Probability functions are now defined on a *σ-field* (*σ-algebra*)—a field $\mathscr{F}$ that is further closed under *countable* unions.[8] For technical reasons, we do not require $\mathscr{F}$ always to be the power set (the set of *all* subsets) of Ω, although it sometimes is (the power set is always a σ-field). As we will see shortly, it turns out that in some infinite probability spaces, not all subsets of Ω are 'well behaved'.

[6] See especially the chapters in this volume by Kenny Easwaran (9), Terrence Fine (11), James Hawthorne (13), and Aidan Lyon (8) for discussions of unorthodox theories of probability.

[7] We may think of complementation as corresponding to negation ('not'), and union as corresponding to disjunction ('or'). $\mathscr{F}$ is closed under complementation just in case for all $A \in \mathscr{F}$, $\bar{A} \in \mathscr{F}$. $\mathscr{F}$ is closed under union just in case for all $A \in \mathscr{F}$ and $B \in \mathscr{F}$, $A \cup B \in \mathscr{F}$.

[8] See Aidan Lyon's chapter (8) in this volume for discussion of some arguments against the requirement that probabilities be defined on a σ-field, and even that they be defined on a field.

The third axiom is correspondingly strengthened:

3'. (Countable additivity) If $A_1, A_2, A_3, \ldots$ is a countable sequence of (pairwise) disjoint sets, each belonging to $\mathscr{F}$, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n).$$

The status of countable additivity is a topic of ongoing debate—see Aidan Lyon's chapter (8) in this volume.

1.7.2 Conditional probability

Much as before, Kolmogorov then defines the *conditional probability of A given B* by the ratio of unconditional probabilities:

$$P(A|B) = {}^{P(A \cap B)}\!/_{P(B)}, \text{ provided } P(B) > 0.$$

Intuitive though this definition is, it has some ramifications that may be considered problematic. Note that the ratio is undefined if either or both of the unconditional probabilities are undefined, or if $P(B) = 0$. Yet there can be possible events whose probabilities are undefined, so-called *non-measurable sets*. And there can be other possible events whose probabilities are 0—'probability 0 does not imply impossible' as textbooks, and Kolmogorov himself, caution us! For example, let Ω be the [0, 1] interval; let $\mathscr{F}$ be the smallest σ-field that includes all of the open intervals (a, b) in [0, 1]; and let P be the uniform distribution, so that $P([a, b]) = b - a$. We may imagine that we are throwing a dart with an infinitely sharp point at the [0, 1] interval, with an equal probability of the dart landing in any sub-interval of a given size. Although our dart will definitely land on *some* real number x, $P(\{x\}) = x - x = 0$ for all x in [0, 1]. Moreover, it is possible (assuming the axiom of choice) to construct a set $S \subset [0, 1]$ that cannot consistently be assigned *any* probability. (Such a set will be an uncountable set of disjoint numbers, and will not be in the σ-field $\mathscr{F}$.) So Kolmogorov's definition does not guarantee that certain intuitive constraints on conditional probability are met—for example, that the probability of an event, *given itself*, is 1. See Easwaran (this volume, 9) for further discussion.

The law of total probability can now be given an infinitary formulation. (This requires the full strength of countable additivity.) Suppose there is a countably infinite partition of propositions $\{A_1, A_2, \ldots\}$. Then for any B,

$$P(B) = \sum\nolimits_{i=1}^{\infty} P(B|A_i)\, P(A_i).$$

when all of the conditional probabilities are defined. Bayes' theorem can similarly be given an infinitary formulation. Suppose there is a countably infinite partition of hypotheses $\{H_1, H_2, \ldots\}$, and evidence E. Then for each i,

$$P(H_i|E) = \frac{P(E|H_i)\, P(H_i)}{\sum_{j=1}^{\infty} P\left(E|H_j\right) P(H_j)}$$

when all of the conditional probabilities are defined.

1.7.3 Independence

Independence of A and B is defined as before. The factorization formulation

$$P(A \cap B) = P(A)\,P(B)$$

generalizes to any finite number of events. Events $A_1, A_2, \ldots, A_n$ are independent just in case the probability of any intersection of the events factorizes. That is, for any $1 < j \leq n$, and any $1 \leq k_1 < \ldots < k_j \leq n$:

$$P(A_{k_1} \cap \ldots \cap A_{k_j}) = P(A_{k_1}) \ldots P(A_{k_j}) \quad \text{(Factorization)}$$

We can even generalize the factorization formulation to infinite—even uncountable—collections of events. Such a collection of events is independent just in case each of its finite subcollections is.

It is with the notion of independence that we arrive at the part of probability theory that is distinctively *probabilistic*. Axioms 1–3, including 3', are general measure-theoretic axioms that apply equally to length, area, and volume (suitably normalized), and even to mass (suitably understood). The notion of independence, by contrast, we find only in this specifically probabilistic application of measure theory. Kolmogorov (1933/1950) writes: 'Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probability its peculiar stamp …We thus see, in the concept of independence, at least the germ of the peculiar type of problem in probability theory.' (8–9). Note that the elegant definition of independence in terms of factorization is possible because probabilities are values in the [0, 1] interval. Suppose, for example, that probabilities were values in the [0, 100] interval. (After all, it is common to report probabilities in the form of percentages.) The probability that a fair coin lands heads on its first toss would then have to be 50; so too the probability that it lands heads on its second toss. But the probability that it lands heads on both those tosses would presumably have to be $25 \neq 50 \times 50$.

1.8 The many bearers of probabilities

This is the probability theory that undergraduates in mathematical statistics are taught, and it is the cornerstone of most applications of probability in the sciences. Note that a specific choice about the bearers of probability has been made: probabilities attach to *sets*. Philosophers often give this a specific interpretation: the basic elements are *possible worlds*, and sets of them are *propositions*. So on this view, the bearers of probabilities are propositions.

Some philosophers, by contrast, regard probabilities as attaching to *sentences* in some formal language. Typically they impose axioms on a probability function P defined on these sentences, which are intended to parallel Kolmogorov's axioms above:

(I) $P(X) \geq 0$ for all sentences X.
(II) $P(T) = 1$ for any tautology (logically necessary sentence) T.

(III) $P(X \vee Y) = P(X) + P(Y)$ if X and Y are incompatible sentences (X & Y is a logical contradiction).

Typically these philosophers either do not raise the question of how these sentential axioms relate to Kolmogorov's set-theoretic axioms, or they dismiss it with some offhand remark to the effect that they are 'equivalent' or 'interchangeable'.

They are not. There are as many ways of understanding the notion of 'tautology' or 'contradiction' as there are distinct logics—classical, intuitionistic, paraconsistent, quantum, etc. (see Priest 2008)—and each corresponds to a different probability theory. Since they are not equivalent to each other, clearly the sentential formulations cannot all be equivalent to the set-theoretic formulation. To be sure, classical logic is typically presupposed. But then note the glaring absence of the sentential counterpart to (3)': where is (III)', telling us that probabilities over sentences are *countably* additive? It is absent for good reason: in classical logic, one cannot even *form* infinitely long sentences, so (III)' cannot even be *stated*. (The ellipsis dots '…' are not a sentential connective!) Hence we see that questions arise for the sentential formulation of probability theory that have no counterpart for its set-theoretic formulation. (For a discussion of probability and non-classical logics, see J. Robert G. Williams' chapter (12) in this volume.)

Going in the other direction, debates about the foundations of set theory presumably should have an impact on the set-theoretic approach to probability, but they may have no impact on the sentential approach. (For example, the proof of the existence of 'non-measurable sets', mentioned above, assumes the axiom of choice, a subject of some controversy.) So much for the received view that the approaches are interchangeable, and that opting for one over another is merely a matter of convenience.

In fact, the sorts of things to which people regard probabilities as attaching are highly varied. Here's a sample (our underlinings):

> A probability measure $P(.)$ …is a numerically valued set function …
>
> Chung (1974).

> We shall develop the probability calculus as it applies to *statements*
>
> Skyrms (2000).

> I think of chance as attaching in the first instance to propositions
>
> Lewis (1980/1986).

> … primary intensions can play the role of objects of credence for Bayesian theory
>
> Chalmers (2011).

> [M]ost cost-benefit analysts who address probabilities appear to hold a frequentistic view in which they are seen as characteristics of events or processes.
>
> Fischhoff et al. (1981).

This list could be continued at length—other authors say that the bearers of probabilities are *event types, sets of states, outcomes, infinite sequences of outcomes, abstract formulae, properties*, and so on. This diversity of accounts of the bearers of probabilities becomes less surprising when one considers the diversity of interpretations of probability itself—see the several chapters on them in this volume.

1.9 Random Variables

Random variables allow us to give more mathematical structure to the events—or what have you—on which probability is defined.

1.9.1 Definitions

Recall that a *probability space* is a triple $< \Omega, \mathscr{F}, P >$, where Ω is a set of outcomes or basic possibilities, and $\mathscr{F}$ is a set of subsets of Ω. A *random variable* is a function $X: \Omega \rightarrow \Re$ that meets further conditions to be described shortly. So it is a function that maps elements of the outcome space Ω to real numbers. The intuitive idea is that a random variable identifies the value of some quantity in each element ω of the outcome space. Note that the term 'random variable' is something of a misnomer, since a *function* from one set to another is neither random nor a variable. Even the determination of a random variable's input need not be random. (See Antony Eagle's chapter on 'Probability and Randomness' (21) in this volume for more on how the relationship between the two notions must be handled with care.)

Here is a simple example. Suppose that we are rolling two dice, one red and one white. The outcome space Ω will consist of ordered pairs $< i, j >$, where the first member corresponds to the result of the red die roll, and the second member corresponds to the result of the white die roll. Hence $\Omega = \{< 1, 1 >, < 1, 2 >, \ldots, < 6, 6 >\}$. The random variable W is defined on Ω as follows: $W(< i, j >) = j$. Intuitively, W represents the result of the white die roll. Another random variable is Y, where $Y(< i, j >) = i + j$. Y represents the sum of the two dice (a random variable of interest when playing Monopoly).

Let $R(X)$ be the *range* of the variable X, the set of possible values X can assign. If $x \in A \subseteq R(X)$, we will write '$X = x$' as a shorthand for $\{\omega \in \Omega: X(\omega) = x\}$ and '$X \in A$' for $\{\omega \in \Omega: X(\omega) \in A\}$. That is, $X = x$ is the set of all outcomes in which X takes the value x; $X \in A$ is the set of all outcomes in which X takes a value in A. If $R(X)$ is finite or countably infinite, then we require that $X = x \in \mathscr{F}$ for all x in $R(X)$. This means that $X = x$ is an event that can be assigned a probability. We can think of $X = x$ as being like a proposition, stating the value of X. We want to be able to talk about the probability of such propositions being true. Since $\mathscr{F}$ is closed under countable unions, $X \in A$ will be in $\mathscr{F}$ for arbitrary subsets A of $R(X)$.

If $R(X)$ is uncountable, e.g. if $R(X) = \Re$, then it is necessary to impose a stricter condition. We require that $X \in H$ be in $\mathscr{F}$ for any *Borel set* H. The Borel sets in $\Re$ are the members of the smallest σ-algebra that contains all of the open intervals in $\Re$. This will include all of the open intervals, closed intervals, half-open intervals, and singleton sets, as well as their complements and countable unions. In short, it will include all the 'nice' sets one is likely to encounter. However, it will not include just any arbitrary collection of real numbers.

The definition of a random variable can be broadened to allow for functions with other kinds of range. For example, there can be vector-valued random variables, which have ranges in $\Re^k$ for some k. Further extensions are possible as long as the range of each variable has a topological structure that is analogous to the structure of the Borel sets in $\Re$. We will restrict our attention to real-valued random variables in what follows.

Every event has an associated random variable. If $A \in \mathscr{F}$, the *characteristic function*[9] or *indicator function* for A is a random variable X_A such that $X_A(\omega) = 1$ if $\omega \in A$, and $X_A(\omega) = 0$ otherwise.

1.9.2 Independence

Probabilistic independence can be defined for random variables much as it is for events. Random variables $X_1,\ldots, X_n$ are independent just in case

$$P(X_1 \in H_1 \cap \ldots \cap X_n \in H_n) = P(X_1 \in H_1)P(X_2 \in H_2)\ldots P(X_n \in H_n)$$

whenever $P(X_i \in H_i)$ is defined for all i. For an infinite sequence X_1, X_2, …, of random variables to be independent, this equation must hold for each n.

1.9.3 Distributions and Density Functions

1.9.3.1 *Distributions*

If X is a random variable, the *distribution* of X is a function μ_X defined on subsets of the range of X. The distribution corresponds to the probability that the value of X will fall within a certain range. That is:

$$\mu_X(H) = P(X \in H)$$

whenever the latter quantity is defined. μ_X will be a probability function whose outcome space is the range of X. For example, consider our earlier illustration involving the roll of two dice. The outcome space consists of ordered pairs of the form $< i, j >$, and we considered a random variable $W(< i, j >) = j$, representing the number resulting from the white die roll. If the die is fair, then the distribution of this variable will be $\mu_W(\{j\}) = 1/6$ for $j = 1, 2,\ldots, 6$.

1.9.3.2 *Identical Distributions*

If two random variables X and Y have the same distribution, then they are *identically distributed*. Returning to our running example, suppose both dice are fair, and let $R(< i, j >) = i$, so that R corresponds to the result of the red die roll. R and W are distinct random variables: for example, $R(< 2, 3 >) = 2$, while $W(< 2, 3 >) = 3$. But R and W are identically distributed, since $\mu_R(\{i\}) = \mu_W(\{i\}) = 1/6$ for $i = 1, 2,\ldots, 6$.

1.9.3.3 *Joint Distributions*

Given at least two random variables defined on the same probability space, their *joint distribution* gives the probability that each variable's value falls in a certain range specified

[9] In mathematical statistics the term 'characteristic function' is sometimes used with a distinct meaning.

for that variable. For two random variables X and Y,

$$\mu_{X,Y}(H_1, H_2) = P(X \in H_1 \cap Y \in H_2).$$

This can be extended to any finite number of random variables in the obvious way.

Continuing our dice example, the joint distribution of R and W is given by $\mu_{R,W}(\{< i, j >\}) = 1/36$ for $i = 1, 2, \ldots, 6$ and $j = 1, 2, \ldots, 6$.

1.9.3.4 *The Binomial Distribution*

One distribution that arises frequently is the *binomial distribution*. Suppose we have a sequence of random variables $X_1, \ldots, X_n$ such that each variable has range $\{0, 1\}$, and the variables are independent and identically distributed with $P(X_i = 1) = p$. Let S be a new random variable defined by $S(\omega) = \Sigma_{i=1,\ldots,n} X_i(\omega)$. Think of 1 as a designated outcome, and S as the number of times this outcome appears in n trials. For example, suppose that we have a coin with probability p of landing heads when tossed. We toss the coin n times; then X_i could represent the outcome of the ith toss, with $X_i = 1$ corresponding to heads and $X_i = 0$ corresponding to tails. S is the total number of heads.

The distribution of S is the binomial distribution:

$$\mu_S(m) = \binom{n}{m} p^m (1-p)^{n-m}$$

where $\binom{n}{m}$ (read 'n choose m') is equal to $n!/m!(n - m)!$. ($n!$, read 'n factorial', is the product $1 \times 2 \times \ldots \times (n-1) \times n$.) $p^m(1-p)^{n-m}$ is the probability of getting any particular sequence of m 1's and $n - m$ 0's, while $\binom{n}{m}$ is the number of distinct such sequences.

1.9.3.5 *Density Functions*

If the range of X is finite or denumerable, the distribution of X is fully determined by probabilities of the form $P(X = x)$. For any other set H, $P(X \in H)$ is determined by summing over the elements of H: $P(X \in H) = \Sigma_{x \in H} P(X = x)$.[10] However, if the range of X is $\Re$ or some other uncountable set, this is not possible. Since X can take uncountably many values, we must have $P(X = x) = 0$ for almost all x.[11] Thus merely specifying all the values of $P(X = x)$ will tell us very little about the distribution of X. In this case, it is often convenient to represent the distribution of X in terms of a *density function*. A density function for X is a function ρ_X on the range of X such that whenever $P(X \in H)$ is defined:

$$P(X \in H) = \int_H \rho_X(x)dx.$$[12]

[10] If H is denumerable, countable additivity (as opposed to merely finite additivity) will be required.

[11] Specifically, $P(X = x)$ can be positive for at most countably many values of x.

[12] The case in which P assigns positive weight to specific values $X = x$ is somewhat complicated, and we will not address it here.

Probabilistic density behaves much like the more familiar notion of mass density. We can represent the distribution of mass in a rock by a density function. (This is an idealization, since the rock is composed of atoms. But let us imagine that the rock is made of continuous stuff.) If the rock is not homogeneous, the mass density will be higher in some places than others. In assigning a mass density to a point, we are not saying that that point, by itself, has any positive mass. Nor can we calculate the mass of the rock by just 'adding up' the densities at every point. Integration is the analogue of addition when we have an uncountable number of quantities to sum.

Here is a very simple example. Suppose that X assigns values in $[0, 1]$. X is *uniformly distributed* on $[0, 1]$ just in case whenever $0 \leq a \leq b \leq 1$, $P(X \in [a, b]) = b - a$. For example, the probability that X is between 5/16 and 6/16 is the same as the probability that X is between 14/16 and 15/16, namely 1/16. Then the simplest density function for X is $\rho_X(x) = 1$ for $0 \leq x \leq 1$, and 0 otherwise. That, e.g., $\rho_X(5/16) = 1$ does not mean that $P(X = 5/16) = 1$; indeed $P(X = 5/16) = \int_{5/16}^{5/16} 1\, dx = 0$.

We may define joint densities of two or more continuous random variables analogously to the way we defined joint distributions for two or more discrete random variables. We may also formulate continuous versions of the law of total probability and of Bayes' theorem, using density functions.

1.9.3.6 *The Normal Distribution*

An important distribution in probability and statistics, and frequently appealed to in the natural and social sciences, is the *normal distribution*, or *Gaussian distribution*, otherwise called 'the bell curve'. This is actually a family of distributions, defined by two parameters m and σ, corresponding to the *mean* or *expectation*, and the *standard deviation*, respectively.[13] We will define these quantities in the next section, as well as the *variance* σ^2. We will write $N(m, \sigma^2)$ for the normal distribution with mean m and variance σ^2. The density for a normal distribution has the form:

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}$$

In the simplest case where m is 0 and σ is 1, this simplifies to:

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

1.9.3.7 *Cumulative Distribution Functions*

The *cumulative distribution function* of a real-valued random variable X is a function F_X such that:

$$F_X(x) = P(X \leq x) = \mu_X((-\infty, x]).$$

If X has density function ρ_X, then $F_X(x) = \int_{-\infty}^{x} \rho_X(z)dz$.

[13] The Greek letter 'μ' is often used for the mean, but we will avoid this since we are using it for the distribution of a variable. We hope that use of 'σ' for standard deviation, and for σ-fields will not cause confusion, since these arise in different contexts.

We can define *joint cumulative distribution functions* for two or more random variables in natural ways. For example, for the random variables X and Y it is

$$F_{X,Y}(x,y) = P(X \leq x \cap Y \leq y) = \mu_{X,Y}((-\infty,x],(-\infty,y]).$$

1.9.3.8 *Convergence in Distribution*

The cumulative distribution function is used to define a special type of convergence called *convergence in distribution*. Suppose that random variables X, X_1, X_2,… have cumulative distribution functions F, F_1, F_2,…, respectively. Then the sequence X_1, X_2,… *converges in distribution* to X just in case

$$\lim_{n\to\infty} F_n(x) = F(x)$$

for every x at which F is continuous. (Since F is non-decreasing, and takes values between 0 and 1, it can have at most countably many 'jumps', and will be continuous everywhere else.) We will see some important examples of convergence in distribution in the section on limit theorems.

1.9.4 Expectation

The *expectation*, or *expected value*, of a random variable generalizes the familiar concept of an arithmetic mean or weighted average. James Rodriguez of Colombia scored six goals in five matches during the 2014 football (soccer) World Cup, scoring one goal in each of four matches, and two goals in a fifth match. His average was 1.2 goals per match: six goals divided by five matches. If a student scores 87, 79, and 82 on three tests, and the tests are worth 20%, 30%, and 50% of the final grade (respectively), then the student's grade will be $(.2 \times 87) + (.3 \times 79) + (.5 \times 82)$ or 82.1.

If X is a random variable whose range is finite or countably infinite, the expectation of X is defined by

$$E(X) = \sum_{x\in R(X)} xP(X = x).$$

This is the result of summing all the possible values of X, weighting each one by the probability that X takes this value. We can also express the expectation of X using its distribution μ_X:

$$E(X) = \sum_{x\in R(X)} x\mu_X(\{x\}).$$

Note that the terms 'expectation' and 'expected value' of a random variable are misleading, suggesting as they do that the value is somehow probable or to be predicted. Indeed, the expectation of a random variable is often not even a possible value for it. Nobody *expected* Rodriguez to score 1.2 goals in a particular match!

A useful identity is that the expectation of the characteristic function for an event is the probability of the event:

$$E(X_A) = P(A).$$

Expectations play a central role in *decision theory* (see Lara Buchak's chapter (36) in this volume). The expectation of a gamble—and more generally, of an action—is supposed to codify how desirable it is. For example, suppose that one thousand tickets are sold in a lottery. Ten tickets will win a prize of $100, and one ticket will win a prize of $1000. The expected winnings of a single ticket are [$100 ×(10/1000)] + [$1000 ×(1/1000)] or $2. Thus, (subject to a number of assumptions) $2 would be a fair price for the ticket.

If $R(X)$ is uncountable, the expectation must be defined in terms of the density function ρ_X:

$$E(X) = \int_{R(X)} x\rho_X(x)\,dx$$

Instead of adding up the values of X and weighting them by discrete probability values, we integrate over the possible values of X, weighting them by the density function evaluated at those values.

1.9.5 The St. Petersburg Paradox

The St. Petersburg Paradox is a paradox involving expectation, first published in the 18th century in the *Commentaries of the Imperial Academy of Science of St. Petersburg*. Suppose you are offered the chance to play the following game: A fair coin will be flipped as many times as necessary before it lands heads. If it lands heads on the first toss, you will win $1; if it lands heads on the second toss, you will win $2; on the third toss, $4; and so on, doubling each time.[14] The expected payoff of the game is

$$\left(1 \times \frac{1}{2}\right) + \left(2 \times \frac{1}{4}\right) + \left(4 \times \frac{1}{8}\right) + \ldots = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} \ldots = \infty.$$

This is very counterintuitive, since the game is certain to pay only a finite amount. It would not seem rational to pay an infinite amount of money to play this game.

A variety of solutions to the paradox have been proposed. We will not canvass these solutions here. The calculation of the expectation is mathematically correct. This demonstrates that if the random variable X is unbounded (there is no finite number n such that $|X|$ is always less than n), then $E(X)$ can be infinite.

1.9.6 Variance and Standard Deviation

The *variance* and *standard deviation* of a random variable are measures of how spread out the variable is. They measure the range of values that a variable can take, weighted by the probability that the variable takes those values. If $E(X) = m$, then the variance of X, σ_X^2 is defined by:

$$\sigma_X^2 = E\left((X-m)^2\right).$$

[14] In the event that the coin lands tails forever (which has probability zero), there is no payoff.

The variance is the expectation of X's deviation from its mean value, where the size of the deviation is measured by $(X - m)^2$. By squaring the difference, the deviation from the mean will count as positive, regardless of whether X is above the mean or below it. Moreover, larger deviations will count for more than smaller deviations. The standard deviation of X, σ_X, is the square root of the variance.

Suppose that a coin has probability p of landing heads. We represent a toss of the coin by a random variable X, with $P(X = 1) = p$, and $P(X = 0) = 1 - p$. $E(X) = p$, so the variance of X will be $E((X - p)^2) = (1 - p)^2\, p + p^2(1 - p) = p(1-p)$.

For a second example, suppose that X takes values in $[0, 1]$ with uniform distribution. Then $E(X) = 0.5$, and $\sigma_X^2 = \int_0^1 (x - 0.5)^2\, dx = 1/12$.

1.10 Limit Theorems

There are a number of important theorems that describe the limiting behaviour of infinite sequences of random variables, X_1, X_2,…We will discuss only a few of the best-known versions here. Hans Fischer's chapter (5) in this volume discusses the history of many of these theorems; see also Terrence Fine's chapter (11) for further discussion of them, and relatives of them.

1.10.1 The Laws of Large Numbers

The Laws of Large Numbers formalize what is popularly known as 'the law of averages'—roughly, the tendency in the long run for random processes to yield results in proportion to their probabilities.[15] Let $X_1, X_2,\ldots$be a sequence of independent and identically distributed variables, each having finite expectation m. Let $S_n = X_1 + X_2 + \ldots + X_n$, so that the average value of the first n variables is S_n/n. The *Weak* and *Strong Laws of Large Numbers* characterize different senses in which the average value S_n/n converges to m.

The Weak Law of Large Numbers states that:

$$\text{For every } \varepsilon > 0,\ \lim_{n\to\infty} P(|(S_n/n) - m| < \varepsilon) = 1.$$

That is, the probability that the relative frequency S_n/n differs from m by less than a fixed ε converges to 1, for every $\varepsilon > 0$. This implies that the sequence $\{S_n/n\}$ converges in distribution to a variable that always takes the value m.

The Strong Law of Large Numbers states that:

$$P\left(\lim_{n\to\infty} (S_n/n) = m\right) = 1.$$

[15] A common mistake is to think that random processes tend to regulate themselves, so as to display this tendency even in the short run. This is called *the gambler's fallacy*. For example, after seeing a sequence of coin tosses yield a run of heads, people often feel inclined to bet on tails, thinking that somehow it is 'due'. This is nonsense if the trials are independent.

The sequence $\{S_n/n\}$ converges to m everywhere in the outcome space Ω, except possibly on a set with probability zero.

For example, suppose that the variables X_1, X_2,…represent repeated tosses of a coin, with probability p of landing 'heads' ($0 < p < 1$). As usual, we let $X_i = 1$ represent 'heads' on the ith toss, and $X_i = 0$ represent 'tails'. Then $m = p$; the expected value of each X_i is just the probability of 'heads'. S_n is the number of heads among the first n tosses, and S_n/n is the relative frequency of heads among the first n tosses. The Strong Law of Large Numbers says that with probability 1, the frequency S_n/n will converge to p as n increases. We have noted earlier that probability 0 does not imply impossibility; the flip side to this is that probability 1 does not imply necessity. And indeed, the convergence here is not necessary. There are some (in fact, infinitely many) sequences of outcomes in which the proportion of heads will not converge to p. For example, the sequence in which every flip lands heads is one possibility represented in the outcome space. But the set of all such sequences must be assigned a probability of 0.

In slogan form: the Strong Law of Large Numbers is the happy claim that there is *probability 1 of convergence* of the relative frequency to the true probability. The Weak Law is the happy claim about the *convergence of probability to 1* of the relative frequency's discrepancy from the true probability being arbitrarily small. The former is called *almost sure convergence,* the latter *convergence in probability*.

As the names suggest, the Strong Law implies the Weak Law, but not vice versa. An important difference between the two is that the Weak Law involves assignments of probabilities to finitary events only. That is, it concerns the probability that S_n/n takes particular values, for finite values of n. The Strong Law assigns a probability to an event that is infinite in character; it concerns the probability of a single event involving all of the S_n at once. The Weak Law is much simpler to prove, and it requires only that the probability function P be *finitely* additive. The Strong Law requires that P be countably additive.

The Laws of Large Numbers are often cited as explaining why we tend to see long-run relative frequencies that approximate what we take to be the corresponding probabilities. For example, in 100 tosses of a fair coin, we typically see roughly 50 heads and 50 tails, and thus a relative frequency of heads of roughly $50/100 = ½$. However, the Laws are *limit* theorems, and as such they have no implications for what we will see in a given finite number of trials (such as 100), nor even for what we will *probably* see. The Weak Law tells us that a certain sequence of probabilities converges to 1. But for all it says, this may happen very slowly. (Compare: the sequence $a_n = 1 - n^{-1/100}$ converges to 1, but its 100th term is roughly 0.05, which is nowhere near 1. The limiting behaviour of the sequence is a very poor guide to what happens early in the sequence.) The Strong Law tells us that the *limiting* relative frequency is almost surely equal to the true probability. But for all it says, a relative frequency *after a given finite number of trials* may be wildly different from this probability.

In the special case of a variable with two possible values, it is the *binomial distribution* that tells us how close we can expect the relative frequency to be to the mean. In our example, the number of heads in 100 tosses of a fair coin has a binomial distribution with $n = 100$ and $p = ½$. The probability that the number of heads is between 45 and 55 (inclusive), and thus that the relative frequency is within 0.05 of 1/2, is about 0.73, which is rather high. As n increases, this probability increases rather quickly.

1.10.2 The Central Limit Theorem

As above, assume that $X_1, X_2, \ldots$ is a sequence of independent and identically distributed variables with finite expectation m. We will now assume, in addition, that the variables also have finite variance σ^2. S_n is defined as above. The Laws of Large Numbers tell us that S_n/n converges to m. The *Central Limit Theorem* tells us in more detail what the distribution of values of S_n/n about m looks like. Specifically, it tells us that

$$\sqrt{n}\,((S_n/n) - m) \text{ converges in distribution to } N(0, \sigma^2).$$

No matter what the original distribution of the X_i's, as you add more of them together, the distribution converges to a normal distribution. Specifically, if you take the amount by which S_n/n differs from m, and multiply it by $\sqrt{n}$, the distribution of the resulting quantity approaches a normal distribution.

The conditions required for a sequence of distributions to converge to a normal distribution can be weakened in a variety of ways. These theorems are often thought to explain why distributions found in nature often approximate normal distributions. The usual explanation is that this is to be expected whenever a variable is the sum of many small, independent components. However, see Lyon (2014) for concerns about this type of explanation. Instead, notice in our coin example that the binomial distribution with 100 trials and probability of ½ closely approximates a normal distribution with mean 50 and variance 25.

1.11 Conclusion

The essentials of probability theory are now before you. The first pass suffices for most philosophical applications. Despite its simplicity, the paradoxes show that even elementary probability theory must be handled with some care, as it apparently does not come entirely naturally to us. (This should not come as a surprise: we are susceptible to various fallacies in reasoning even when the underlying logic is simple—for example, affirming the consequent.) The second pass provides rigorous foundations for more advanced applications. It also provides further tools that can be used to clarify and to sharpen philosophical discussions.

You should now be well placed to read the state of the art in philosophy of probability, and various applications of probability in philosophy, covered in the rest of this handbook. Enjoy!

References

Berkson, J. (1946) Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biometrics Bulletin*. 2. 3. pp. 47–53.

Bertrand, J. L. F. (1889) *Calcul des Probabilités*. Paris: Gauthier-Villars et fils.

Chalmers, D. (2011) Frege's Puzzle and the Objects of Credence. *Mind*. 120. pp. 587–635.

Chung, K. L. (1974) *A Course in Probability Theory*. London: Academic Press.

Fischhoff, B., Lichtenstein, S., Slovic, P., Derby, S., and Keeney, R. (1981) *Acceptable Risk*. Cambridge: Cambridge University Press.

Fitelson, B. and Hájek, A. (2014) Declarations of Independence. *Synthese*. 10.1007/s11229-014-0559-2. [Online] Available from: http://fitelson.org/doi_pub.pdf.

Gardner, M. (1959) Mathematical Games. *Scientific American*. October. pp. 180–2.

Glymour, C. (2001) *The Mind's Arrows*. Cambridge, MA: MIT. Press.

Kolmogorov, A. N. (1933/1950) *Foundations of the Theory of Probability*. Translated from the German. London: Chelsea Publishing Company.

Lewis, D. (1980/1986) A Subjectivist's Guide to Objective Chance. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability*. Vol. II. Berkeley, CA: University of California Press. (Reprinted in *Philosophical Papers*. Vol. II. Oxford: Oxford University Press.)

Lyon, A. (2014) Why are Normal Distributions Normal? *The British Journal for the Philosophy of Science*. 65. 3. pp. 621–49.

Pearl, J. (2009) *Causality: Models, Reasoning, and Inference*. 2nd ed. Cambridge: Cambridge University Press.

Priest, G. (2008) *An Introduction to Non-Classical Logic*. 2nd ed. Cambridge: Cambridge University Press.

Simpson, E. (1951) The Interpretation of Interaction in Contingency Tables. *Journal of the Royal Statistical Society*. Series B. 13. pp. 238–41.

Skyrms, B. (2000) *Choice and Chance*. 4th ed. Belmont, CA: Wadsworth.

Van Fraassen, B. (1989) *Laws and Symmetry*. Oxford: Clarendon Press.

Westbrooke, I. (1998) Simpson's Paradox: An Example in a New Zealand Survey of Jury Composition. *Chance*. 11. 2. pp. 40–2.

Yule, G. (1903) Notes on the Theory of Association of Attributes in Statistics. *Biometrika*. 2. pp. 121–34.

PART I

HISTORY

CHAPTER 2

PRE-HISTORY OF PROBABILITY

JAMES FRANKLIN

It has been common to begin histories of probability with the calculations of Fermat and Pascal on games of chance in 1654. Those calculations were the first substantial mathematical results on stochastic phenomena, which, according to a frequentist philosophy of probability, are the true subject matter of the field. But there is great philosophical interest in studying earlier ideas. That is because we see a struggle to understand a range of basic concepts about probability and uncertain evidence, before the straightjacket of a single formalization was imposed. There is still much to learn from seeing how the problems now studied with the aid of probability theory were dealt with in "bare hands" fashion before that formalization was available.

2.1 Two Concepts: Logical/Epistemic Versus Factual/Stochastic Probability

Certain philosophical distinctions are needed in order to identify what body of ideas in ancient and early modern texts should be regarded as probability. The main distinction required is that between factual or aleatory probability, on the one hand, and logical or epistemic probability on the other (introductions in Hacking 2006: ch. 2; Mellor 2005: ch. 1; Franklin 2009: ch. 10; finer distinctions in Hájek 2002/2011). *Factual* or *stochastic* or *aleatory* probability deals with chance set-ups such as dice-throwing and coin-tossing, which produce characteristic random or patternless sequences of outcomes. The calculus of probability applies straightforwardly to it.

By contrast, *logical* or *epistemic* probability is concerned with how well-supported a conclusion is by a body of evidence. A concept of logical probability is employed when one says that, on present evidence, the steady-state theory of the universe is less probable than the big bang theory, or that an accused's guilt is "proved beyond reasonable doubt" though not absolutely certain. How probable a hypothesis is, on given evidence, determines the degree of belief it is rational to have in that hypothesis, if that is all the evidence one has that

is relevant to it. Views of the nature of such probability range from the objective versions of Keynes (1921) and Jaynes (2003), holding that the relation of evidence to hypothesis is a matter of strict logic and so a kind of partial implication, to the views of the subjective Bayesian school (e.g. Earman 1992), according to which there are only degrees of belief in propositions, constrained by the laws of probability.

On any of those views, it is unclear how far epistemic probability can be quantified. As an idealization, it is usual to formalize it as a number P($h|e$) between 0 and 1, expressing the degree to which evidence e supports hypothesis h, subject to the usual laws of conditional probability. But it is controversial whether there is such a precise number for arbitrary e and h. The controversy assumes significance for the prehistory of probability, since most of pre-Pascalian writing that can be considered to be about probability dealt with uncertain evidence, and failure to apply numbers to the evaluation of evidence by early writers is not necessarily a sign of a defect on their part.

2.2 The Law of Evidence and Half-proof

Law is a discipline strongly concerned with continuity maintained through a written tradition. Analysis of concepts is central to it. And the complexities involved in real cases keep the concepts grounded in reality and encourage their development. As a result, law was the matrix in which most development of the concepts of probability took place, both those concepts connected with the law of evidence and those involved in aleatory contracts (contracts whose fulfilment depends on chance, such as insurance, lotteries, and games of chance).

Ancient Greek law was less constrained than later legal systems by theory and precedent, more a rhetorical free-for-all. Therefore skill in constructing "likely" arguments was at a premium. Aristotle gives an example from Corax's *Art of Rhetoric* (c. 460BC): "If the accused is not open to the charge, for instance if a weakling is being tried for assault, the defence is that he was not likely to do such a thing. But if he is open to the charge – if he is a strong man – the defence is still that he was not likely to do such a thing, since he could be sure that people would think he *was* likely to do it." (The word translated as "likely" is *eikos*, literally "like".) (Aristotle, *Rhetoric* 1402a; Franklin 2001: p. 105). Many such arguments are found in the Athenian orators in forensic speeches, some of them based on genuinely likely explanations of the facts rather than shallow appeals to plausibility. Aristotle's *Rhetoric* constructs a theory of arguments applicable to non-necessary matter, and distinguishes between arguments that happen to persuade and those that ought to persuade. In non-necessary matters, one must use likelihoods (*eikota*) and signs, the likely (*eikos*) being what usually happens. There are also arguments from fallible signs and arguments from examples. Non-deductive arguments that ought to persuade are said to be the subject of the science of dialectic, but Aristotle does not develop that topic.

These ideas were taken over into Latin by writers on rhetoric such as Cicero and Quintilian. Cicero writes "That is probable which for the most part usually happens or which is the general opinion or which has in itself some likeness to these." (Cicero: I.46; Franklin 2001: p. 116) However the real Roman contribution was not to rhetoric but to systematic law.

Law deals with evidence, hence with the strength and conflict of evidence. The two great ancient systematizations of law, the Jewish *Talmud* and the Roman *Corpus of Civil Law*, both insisted on very high standards of proof for criminal conviction, but recognized that the standard was short of absolute certainty. In Jewish law, conviction requires two witnesses, and the witnesses must give evidence that is not "based on conjecture" but direct. Roman law holds that conviction may not be "on suspicion", for "It is better to permit the crime of a guilty person to go unpunished than to condemn one who is innocent." The charge must be "supported by suitable witnesses, or established by the most open documents, or brought to proof by signs that are indubitable or clearer than light." (Franklin 2001: pp. 5–7) Torture was a feature of Roman law; judges investigating public crimes should not however begin with torture but use whatever "likely and probable arguments" *(argumentis verisimilibus probabilibusque)* were already available. The word "likely" (*verisimilis),* found earlier in Cicero, is used a number of times in the *Corpus of Civil Law* with very much the common meaning it has in modern English, in examples like "it is very likely *(verisimilius)* that the testator had intended rather to point out to his heirs where they could readily obtain forty aurei ... than to have inserted a condition in a trust." (Franklin 2001: pp. 8–9) Both Jewish and Roman law include a concept of "presumption": a proposition taken to be true until the opposite is proved; some but not all of these are likely but uncertain generalizations. (Franklin 2001: pp. 6, 9; Rabinovitch 1973: pp. 112–15).

Ancient Roman law is rather untheoretical, but when the *Corpus* was rediscovered in Italy in the late eleventh century and began to inform Western medieval law, the commentators on it sought to understand the principles and justifications of the Roman rules. In commenting on evidence, they came to realize that there must be pieces of evidence and presumptions of different strengths. The school of Glossators, of the late twelfth century, united several scattered remarks of Roman law under the classification half-proof *(semiplena probatio)*: "It would seem that ... either the plaintiff proves, or not ... I reply that although according to Aristotle it would seem to be an exhaustive division, it is not so according to the laws. There is a medium, namely, half-proof. Say therefore that in such a case proof has been less than full. Then the judge gives the right to complete the case by oath to the plaintiff ... Full proof is by two witnesses, therefore half proof by one." (Azo of Bologna 1966: pp. 254; Franklin 2001: p. 18) There was never any finer grading attempted, such as quarter-proofs. However, a qualitative grading of the strength of presumptions became standard. In a decree of Pope Innocent III in 1209, the question concerns someone in doubt whether there is some legal impediment to his marriage, and hence wondering if sex with his/her spouse is permitted. It depends on the strength of the evidence:

> It should be distinguished, whether the spouse knows for certain the impediment to the marriage, and then he may not engage in carnal intercourse ... or whether he does not know the impediment for certain, but only believes it ... In the second case, we distinguish, whether his conscience is thus from a light and rash belief, or a probable and discreet one (*ex credulitate levi et temeraria, an probabili et discreta*) ... when his conscience presses his mind with a probable and discreet belief, but not an evident and manifest one, he may render the marriage debt, but ought not to demand it.
>
> (*Corpus Iuris Canonici*: 2, col. 908; Franklin 2001: p. 23)

The theory was elaborated by legal writers and applications made. Torture, reintroduced under the influence of Roman law, could be applied only when there was half-proof of

guilt (to produce the other half). The smallest quantum of evidence considered was the "indiciolum", a diminutive of "indication" or "sign". If in a murder case a witness testifies that the brother of the accused was an enemy of the deceased, "an indication is certainly not thereby proved, although it makes an *indiciolum* which is very small, and of no strength." (Franklin 2001: p. 32). The theory of grades of evidence was laid out with many distinctions in the fourteenth-century commentary of Baldus de Ubaldis, which remained standard for centuries. Massive sixteenth-century expansions weighed down by many examples included Mascardus's *On Proofs* (1584) in three volumes and Menochius's *On Presumptions, Conjectures, Signs and Indications* (1587), in two (Franklin 2001: pp. 28–33, 43–6). Although England had a separate legal tradition from the Continent, Latin concepts such as violent presumptions and half-proofs are sometimes found in seventeenth-century English law. The extraordinary continuity of legal thought is shown in the fact that nineteenth-century American texts on evidence law still quote Menochius and Mascardus (Franklin 2001: pp. 60–1, 44).

Law is central to the story of early modern ideas in ways now hard to appreciate. Besides the preservation of ideas within law itself, legal ideas had a much wider currency than they do today. The originators of mathematical probability were all either professional lawyers (Fermat, Huygens, de Witt) or the sons of lawyers (Cardano and Pascal), and so had some contact with at least the broad concepts of legal thought. Such legal connections were not unusual among intellectual figures. Bacon and Copernicus were also lawyers, Montaigne a judge, Valla a notary, Machiavelli and Arnauld the sons of lawyers, and Petrarch, Rabelais, Luther, Calvin, Donne, Descartes, and Leibniz former law students (Franklin 2001: pp. 348–52).

2.3 The Doctrine of Probabilism in Moral Theology

The Catholic practice of confession led to the development of casuistry, the detailed consideration of "cases of conscience", moral dilemmas on which confessors needed to give advice. The confessional was thought of as something like a miniature court of canon law, and a legal way of thinking was applied to achieving consistency in practical advice (Jonsen and Toulmin 1988). A much-discussed issue in late medieval casuistry was the duty of the conscience in a state of doubt. When one is in doubt about a case, for example when it is unclear whether a rule applies, or when authorities disagree, what should one do? Pope Innocent III laid down an axiom, "In doubtful matters the safer path is to be chosen"; that would be a harsh saying if taken literally, as trivial and over-scrupulous doubts would prevent action.

To distinguish between absolute certainty and the kind of certainty possible in matters of action, the phrase "moral certainty" came to be used. It was introduced by Jean Gerson about 1400, on the basis of a saying of Aristotle that ethics is not a precise science and one must be content with the degree of certainty that the subject admits. One should have, Gerson says, moral certainty that a proposed action is right. To acquire moral certainty, one considers

what usually happens, what authorities say, and what one's own learning suggests (Deman 1936: cols 442–3; Franklin 2001: pp. 69–70).

The discussions on the degree of doubt necessary to excuse one from following a doubtful rule are summarized in Sylvester Prierias's standard manual for confessors (about 1515). The entry for "Probable" is:

> "Probable" is used in two ways . . . Second, as what pertains to opinion. And this in two ways. First, the object of a believed opinion; thus Aristotle (*Topics* I) says that the probable is what seems to be to all, or most, or the wise ... according to the Chancellor [Gerson], what is thus probable is called morally certain ... it is equally vicious for the mathematician to seek the persuasive as for the moralist the demonstrative.
>
> (Sylvester Prierias 1515, quoted in Franklin 2001: p. 71)

Some writers argued for strictness, for example Pope Adrian VI held that a soldier in doubt as to the justice of his king's war ought not to serve in it. But the general tendency was more lax, taking relatively weak doubts to be sufficient to excuse. This tendency reached a culmination in the celebrated doctrine of probabilism, enunciated by the Spanish Dominican Bartolomé de Medina in 1577. According to that doctrine, soon widely accepted in Catholic and to some extent Anglican moral theology, one might follow a course of action that is probable, even if the opposite is more probable. Medina explains what the meaning of the probable is, in terms of having arguments and authorities in favour, in such a way as to allow an opinion to count as probable even when the opposite is more probable:

> Opinions are of two kinds: those are probable which are confirmed by great arguments and the authority of the wise (such as that one may charge interest for delayed payment), but others are completely improbable, which are supported neither by arguments nor the authority of many (such as that one may hold a plurality of benefices) . . . an opinion is not called probable because there are apparent reasons adduced in its favour, and it has assertors and defenders (for then all errors would be probable opinions), but that opinion is probable which wise men assert, and very good arguments confirm, to follow which is nothing improbable. This is the definition of Aristotle . . . It could be argued against this that it is indeed in conformity with right reason, but, since the more probable opinion is more in conformity and safer, we are obliged to follow it. Against this is the argument that no-one is obliged to do what is better and more perfect: it is more perfect to be a virgin than a wife, to be a religious than to be rich, but no-one is obliged to adopt the more perfect of those . . .
>
> (De Blic 1930, translated in Franklin 2001: pp. 74–6; Kantola 1994: pp. 124–30; Maryks 2008: pp. 107–19; Schüssler 2005; Schwartz 2014)

Medina's "probable" is thus close to modern English "arguable". He thinks of "probability" in terms of the reasons (including authorities) in favour of an opinion – in abstraction from the reasons against and the consequent balance of reasons that is normally meant now by the term "probability". The question of balance of reasons is further discussed by the most celebrated Spanish Catholic philosopher, Francisco Suárez. He distinguishes usefully between "positive" and "negative" doubt. A doubt is negative when there are no reasons for either side, positive if there are reasons, but they balance. A soldier who knows nothing about the justice of his king's war may trust the king to know that the war is just (assuming the king is of good reputation). "But if the doubt is positive, and there are probable reasons on either side, I believe they are perhaps obliged to inquire into the truth. If they cannot

arrive at that, they should follow what is more probable, and aid him who more probably is in the right; for when there is doubt on a matter of fact, and one which concerns the harm of our neighbour or the defence of the innocent, one must follow what appears more probable." (Suárez 1944: 2: p. 836; Franklin 2001: pp. 76–9) The distinction is similar to Keynes' concept of the "weight of evidence" for a hypothesis, which may increase as evidence builds up even while its probability, expressing the balance of reasons, stays unchanged (Keynes 1921: ch. 6). Suárez also adds to the conditions for a just war the requirement that there should be "at least a more probable hope of victory, or doubt equally balanced as to defeat or victory, according to the necessity of the state and of the common good." (Suárez 1944: 2: pp. 822–3)

Probabilism became widely accepted in Catholic moral theology, and further developments included a distinction between the "extrinsic" probability of an opinion (the authorities in its favour) and its "intrinsic" probability (the reasons for it). The morally lax consequences of probabilism were taken to extremes in the work of Juan Caramuel Lobkowitz, who was also the author of one of the earliest books on mathematical probability (Fleming 2006). But probabilism, and casuistry in general, were discredited in popular opinion by the savage attack on them in Pascal's *Provincial Letters* (1656). Pascal caricatured the entire Jesuit order, the main target of his attack, as committed to the thesis that the opinion of a single doctor may render an opinion probable and hence allowable. Pascal's prose style is admirable, but there is little honest argument (Franklin 2001: pp. 94–101; Maryks 2008: pp. 127–34). Pascal's polemical success is the ultimate source of the false belief, still found in Hacking's widely-read *The Emergence of Probability*, that the word "probable" before Pascal meant merely "supported by authorities". That is incorrect because "intrinsic" probability, based on reasons, was always part of the understanding of the probability of opinions.

2.4 Evaluating Scientific and Historical Theories

One of the main applications of probabilistic reasoning is to evaluate the strengths of arguments for theories – scientific theories, medical diagnoses, historical claims, and so on. In ancient, medieval, and early modern times, the problems with evaluating uncertain theories were widely recognized and the strength of reasons was discussed, although no coherent and accepted overall theory was developed.

Pre-modern and early modern thought had considerable difficulty with the relation of experimental data to scientific theory, with the result that scientific theories were rarely based solidly on observational evidence. One problem was the deductivist model of science according to which the truths of a science should be theorems derivable from axioms evident to reason. That model of science was promoted in Aristotle's *Posterior Analytics*, realized with success in Euclid's geometry, and revived in a new form by rationalists such as Bacon, Descartes, and Pascal. Another was that the "experience" that scientific theory was supposed to generalize and explain was thought of as commonly known facts, rather than expensively gathered experimental data; in any case, the institutional funding for expensive data collection was rarely available. Particularly notable by its absence is any

attempt to tabulate, summarize, and draw inferences from social science or survey data; the first attempt of that kind was Graunt's analysis of London mortality bills in the 1660s (Franklin 2013).

One science did however have a continuous tradition of observation and its relation to theorizing – astronomy. Ptolemy's summary of ancient Greek and Babylonian astronomy includes a notion similar to the averaging of observations to reduce error: to determine the small time by which the length of the year falls short of $365\,^1/_4$ days, he distributes the total observed difference over many years and argues:

> For the error due to the inaccuracy inherent in even carefully performed observations is, to the senses of the observer, small and approximately the same at any [two] observations, whether these are taken at a large or a small interval. However, this same error, when distributed over a smaller number of years, makes the inaccuracy in the yearly motion [comparatively] greater (and [hence increases] the error accumulated over a long period of time), but when distributed over a larger number of years makes the inaccuracy [comparatively] less. (Ptolemy 1984: pp. 136–7)

Ptolemy also gives some attention to the still-vexed problem of the simplicity of theories. While admitting that in general it is "a good principle to explain the phenomena by the simplest hypotheses possible", he argues that it is different with the heavens because of their divine nature (Ptolemy 1984: pp. 600–1; Franklin 2001: pp. 137–9). Copernicus too worried incessantly about the observational errors in the ancient, medieval and recent data that he used to support his heliocentric theory, but believed in the end that "so much and so great testimony agrees with the mobility of the earth." (Franklin 2001: pp. 146–7)

Both Kepler and Galileo, in arguing for the Copernican hypothesis, speak about the probabilistic relation of evidence to theory with some degree of explicitness. Kepler puts forward the argument that the Copernican theory can do without Ptolemy's complicated system of epicycles, "It is much more probable that there should be some one system of spheres." While geometry is deductive, physics is required when postulating a magnetic-like force moving the planets, so "As is customary in the physical sciences, I mingle the probable and the necessary and draw a plausible *(probabilem)* conclusion from the mixture." Galileo's arguments for Copernicanism include a number of probabilistic ones, including one from the proportionality in the rotations of the spheres:

> … nor do I pretend to draw a necessary proof from this; merely a greater probability (*una maggior probabilità*). The improbability (*l'inverisimile*) is shown for a third time in the relative disruption of the order which we surely see existing among those heavenly bodies whose circulation is not doubtful, but most certain. The order is such that the greater orbits complete their revolutions in longer times, and the lesser in shorter: thus, Saturn … And this very harmonious trend will not be a bit altered if the earth is made to move on itself in twenty-four hours. But if the earth is desired to remain motionless …
>
> (Galileo 1967: pp. 118–9; Franklin 2001: pp. 147–59)

The issues with evidence in biological and social sciences are very different from those in astronomy, because of the variability in the subject matter. That was widely recognized from ancient times. Ancient "sciences" such as divination, astrology, and physiognomics particularly needed excuses for their lack of predictive success, and plenty were forthcoming; Ptolemy (and Kepler), for example, defended astrology as only able to predict tendencies, because of the complexity of the planetary influences on fate (Franklin

2001: p. 165). Medicine was hardly in a better position; as Hippocrates says, "Life is short, art long, opportunity fleeting, experiment deceptive, judgement difficult" (Franklin 2001: p. 166). The ancient conflict between the Dogmatic and Empiric schools of medicine resembles that between Rationalism and Empiricism, with the Empirics defending experience-based knowledge independent of insight into causes ("Is the peasant, until he has learned from one of the philosophers something of the nature and substance of the soil, and what is the nature and substance of rain and wind, and how they come about, unable to know by experience what seeds to sow at certain times and on what soil, if they are to spring and flourish and attain completion and perfection?" Galen 1965: p. 61) The Empirics defended inductive reasoning from what has been observed to happen very many times, and made semi-quantitative distinctions among observed frequencies: "Always, as death in the case of a heart wound; for the most part, as purgation from the use of scammony resin; half the time, as death in the case of a lesion of the dura mater; rarely, as health in the case of a cerebral wound." They also used argument from analogy, with closer similarity justifying more "hope" (Galen 1965: p. 31; Franklin 2001: pp. 168–9). There were few improvements in later times. Islamic, European medieval, and Renaissance medical writers laid down rules of experimentation, in themselves reasonable, but the variability of the subject matter defeated most attempts to achieve knowledge.

Similar problems arise with the evaluation of historical theories. But in addition, the difficulty with evaluating hypotheses in history is that one cannot experiment and usually cannot collect new evidence. Medieval historians, though often credulous by modern standards, recognized problems with the conflict of authorities, the a priori implausibility of stories, and the authenticity of documents, and asked if there might be rules for establishing the credibility of histories. A specially significant case was the Donation of Constantine, a document forged about 750 but widely believed to be authentic, which purported to be a grant by the Emperor Constantine of the entire Western Roman Empire to the Pope, in perpetuity. Lorenzo Valla in the fifteenth century demolished the case for its authenticity, citing anachronisms but principally using an argument from silence: there is no mention of such a remarkable grant in documents of the time or long after. Melchior Cano, in the sixteenth century, laid down some signs of true histories. He recognized the possibility of tension between historical testimony and the antecedent unlikeliness of a claim, which is the conflict at the basis of Hume's argument in *On Miracles*. Cano says that even serious historians like Pliny might be disbelieved if what they say is too incredible, but one should not be too ready with scepticism as it may be due to one's limited experience.

> It would be as if the Mediterranean peoples were to deny the existence of the ocean, or if those who were born on an island in which they saw nothing but hares and foxes should not believe in the lion or the panther, or if, indeed, we should mock at him who speaks of elephants.
>
> (Cano 1734: p. 89; Franklin 2001: pp. 187–94)

2.5 Induction

Contrary to myth, Hume did not discover the problem of induction, though his argument as to why it is unsolvable is original. The problem was one of the many sceptical puzzles posed in ancient times and collected by Sextus Empiricus:

> For, when the Dogmatists attempt to lend credence to a universal by induction from the particulars, in doing this they will consider either all the particulars or only some of them. But if they consider only some, the induction will not be firm, since some of the particulars omitted in the induction may refute the universal; while if they consider all, they will be working at an impossible task, since the particulars are infinite in number and unbounded.
>
> (Sextus Empiricus 2: 204: pp. 160–1; Franklin 2001: p. 203)

Medieval writers of an Aristotelian tendency understood that something was required to fill the logical gap between particular instances and a universal generalization, and hoped that some proposition about causality would accomplish the task. Avicenna, for example, said that if we saw the herb scammony being repeatedly followed by purging, we would argue "we have experienced this often and then reasoned that if it were not owing to the nature of scammony but only by chance, this would happen only on certain occasions." Aquinas held that one might make inductive inferences about the future not only about what happens always, but what has a tendency to happen one way rather than the other: "For example, we can conjecture about future effects depending on free choice by considering men's habits and temperaments, which incline them to one course of action." (Aquinas 1952: p. 378; Franklin 2001: pp. 203–6) Duns Scotus proposed to justify inductive inference "in virtue of this proposition reposing in the soul: 'Whatever occurs in a great many instances by a cause that is not free, is the natural effect of that cause' ... because a cause that does not act freely cannot in most instances produce an effect that is the very opposite of what it is ordained by its form to produce." (Duns Scotus 1304/1962: pp. 109–10; Franklin 2001: p. 206)

The nominalist school, led by William of Ockham, tended to deny such Aristotelian necessary connections, in the light of God's absolute power to produce the opposite of any such alleged necessity. The limit of this tendency was reached in the remarkable work of Nicolas of Autrecourt, the "medieval Hume", soon before the Black Death of 1349. Nicolas argued that no such purported necessities could be reduced to the "necessity of the first principle" – that is, that their denial did not involve a contradiction. A number of Nicolas's propositions were officially condemned by the Church, including one that is a statement of inductive scepticism: "'This consequence is not admitted with any evidence deduced from the first principle: 'Fire is brought near to tow and there is no impediment, therefore the tow will burn.'" (Franklin 2001: pp. 210–16)

Certain Jesuit philosophers of the seventeenth century returned to the problem. Given that, as Ockham had said, God could at any moment suspend the laws of nature, why should one presume that he will not do so? The Jesuit cardinal Juan de Lugo argues:

> As long as the contrary be not proved, we always presume that God without miracles or violence allows secondary causes to act naturally, although indeed God could morally do the opposite . . . Because in case of doubt there suffices a presumption founded on long induction of effects, a bare moral possibility is not sufficient for a prudent doubt or judgement of the opposite. For many things are morally possible, which as long as they are not proved, are not posited, but rather are presumed not to be so . . .
>
> (Juan de Lugo 1868: 2: p. 240; Knebel 1992: p. 189; Franklin 2001: p. 223)

That still does not exactly explain why God should be on the side of the presumption of regularity. The Jesuit Esparza in the 1650s argued that although God could annihilate the

world at any time, the end of the world is *necessarily* a rare event: "it is possible for men to be entirely untouched by any concern for the annihilation of the world, even if God were quite indifferent to that event; because that event belongs to the class of events that can happen but rarely." (Franklin 2001: p. 223; Knebel 1992)

Hobbes insisted on the fallibility of induction, but allowed it to be a good bet. He even offered, if only by way of example, numerical odds for it:

> though a man have always seen the day and night to follow one another hitherto; yet can he not thence conclude they shall do so, or that they have done so eternally: *experience concludeth nothing universally*. If the signs hit twenty times for one missing, a man may lay a wager of twenty to one on the event; but may not conclude it for a truth.
>
> (Hobbes 1839: 4: pp. 17–18; Franklin 2001: p. 224)

If that is the correct odds when there are twenty observations for and one against, it is unclear what the odds should be if there are twenty for and none against. The answer to that is no clearer now than in Hobbes' time.

2.6 Two Argument Forms: "Not by Chance" Arguments and Statistical Syllogisms

Two argument forms that are crucial to modern quantitative probabilistic inference appear in semi-quantitative form in pre-modern writings. One is the "ruling out of the hypothesis of chance", which is essential to modern statistical hypothesis testing. If a "null" hypothesis, say that a drug has no effect, would render improbable the observed evidence of many cures following the drug, the hypothesis is said to be ruled out (at some level of significance). Aristotle argues similarly. Do the stars in their daily revolutions move independently or are they all fixed to a sphere? It is observed that those stars that move in large circles (near the celestial equator) take the same time to rotate as those near the Pole Star which rotate in small circles. Aristotle concludes, "If the arrangement was a chance combination, the coincidence in every case of a greater circle with a swifter movement of the star contained in it is too much to believe. In one or two cases it might not inconceivably fall out so, but to imagine it in every case alike is a mere fiction. Besides, chance has no place in that which is natural . . . " (Aristotle, *On the Heavens:* 289b; Franklin 2001: pp. 133–4; Macdonald 2004). As we saw above, Avicenna proposed to justify induction with such an argument. Similar arguments became a staple of early modern design arguments for the existence of God.

A second argument form is the statistical syllogism (or proportional syllogism or argument from frequencies). If x% of A's are B's, and this is an A, I may conclude with probability x% that this is a B (of course, in the absence of further relevant evidence); if the vast majority of flights that take off land safely, my feeling of safety on takeoff is justified. It is this form of inference that gives rise to the much-discussed reference class problem, arising from the fact that an individual case is a member of many classes A, in which the proportion of B's may differ, hence leading to statistical syllogisms with

conflicting conclusions. Arguments from what happens "for the most part" were common in ancient rhetoric and elsewhere, particularly explicit examples being found in medicine and in Jewish law (Franklin 2001: pp. 169, 172–5; Gabbay and Koppel 2011). Although there were never any attempts at precise quantitative versions of such arguments, there are some remarkable versions in the work of the fourteenth-century mathematical genius Nicole Oresme which involve explicit proportions in infinite sets. In the course of a complex argument against astrology to the effect that the proportion of cases in which it can predict rightly is vanishingly small, Oresme states clearly several times the connection between relative frequency and probability: "It is probable (*verisimile*) that two proposed unknown ratios are incommensurable, because if many unknown ratios are proposed it is most probable that any would be incommensurable with any other":

> if there were some number as to which it were completely unknown what it is or how great it is, and whether it is large or small — as perhaps the number of all the hours that will pass before Antichrist — it will be likely that such a number would not be a cube number. It is similar in games where, if one should inquire whether a hidden number is a cube or not, it is safer to reply that it is not, since that seems more probable and likely (*probabilius et verisimilius*).
>
> (Oresme 1966: pp. 248–51; Franklin 2001: pp. 140–5; Meusnier 2007b)

Unfortunately Oresme's works were little known until recent times.

2.7 Aleatory Contracts: Insurance, Options, Life Annuities, Risks

The business world has always had to deal with uncertainty – ships return or not, debtors repay or default, seasons are good or bad – and those risks are reflected in prices. It is not easy to understand how ancient, medieval, and early modern businesses conceived of risk, since not much was written by businessmen themselves on conceptual matters, and what there is has not been sufficiently studied. However, there is considerable evidence from law and moral theology, which dealt with disputes in business and arguments on the ethics of business dealings.

Ancient Greek and Roman law recognized the maritime loan, a loan advanced for a voyage with repayment only if the ship returned. Variable prices for such loans, for example for differences in seasons, represent an implicit quantification of risk. Roman law says "the price is for the peril." It goes further in accepting that "perils", "hopes", and "expectancies" are entities that can be sold separately and have a price. "Sometimes, indeed, there is held to be a sale even without a thing, as where what is bought is, as it were, a chance (*quasi alea*). This is the case with the purchase of a catch of birds or fish or of largesse showered down. The contract is valid even if nothing results, because it is a purchase of an expectancy (*spei*)." (*Corpus Iuris Civilis, Digest* 18.1.8; Franklin 2001: pp. 259–60) The ancient Jewish law of the Talmud has a similar notion of conditional doubtful claims for which a price may be estimated (Ohrenstein and Gordon 1991). Islamic law, however, forbade any contracts whose fulfilment depended on chance.

In medieval canon law and moral theology, the charging of interest on a loan was condemned as usury. That led to many subtle discussions on what contracts were permissible, for example whether the charging of a fee for late payment might be really a charge for risk instead of genuine interest. It became accepted that prices might reflect common experience of what would probably happen in the future (Piron 2007). Thus Peter Olivi writes, about 1300:

> It is also clear that when someone, by special grace, suffers requisition or sells a crop at a time when it is commonly cheap, but which he firmly intended to keep and sell at a time when it is commonly and probably dearer, he may demand the price which, at the time of requisition, he believed probably would be obtained at the dearer time.
>
> (Olivi 1980: p. 84; Franklin 2001: p. 266; Ceccarelli 1999)

In discussing the amount of compensation for personal injury, he says "the depriver is required to restore only as much as the probability of profit weighs (*quantum ponderat probabilitas talis lucris*)". In another place, concerning transfer of risks for a fee, he writes "the priceable value of a probability *(appreciabilis valor probabilitatis),* or of a probable hope of profit, from the capital is capable of being traded." Such language clearly envisages the quantification of an expectation, in the sense of a potential outcome weighted by the probability of its happening.

At the same period there was discussion of the licitness and pricing of life annuities. It is argued that such contracts can be fair or not, depending on how the price relates to the probabilities of death:

> we see men and women twenty-five years old buying life annuities for a price such that within eight years they will receive their stake back; and although they may live less than those eight years, it is more probable (*probabilius*) that they will live twice that. Thus the buyer has in his favour what happens more frequently and is more probable.
>
> (Alexander [Bonini] of Alexandria 1962: pp. 152–3; Franklin 2001: p. 270)

Fourteenth-century Italy developed insurance contracts, which require the pricing of an expectation. Similar language was used to discuss the price and licitness of insurances (Ceccarelli 2001; Ceccarelli 2007). In the first book on insurance (1488), the Portuguese jurist Santerna explains what the price is for:

> It can be said that the insurer sells only the hope of a future outcome, of which there can well exist a sale . . . from the fact that this hope is uncertain, it might not seem capable of estimation such that in respect of it there could be said to be exceeding of half the just price of its value. But, this is not to be estimated at how much the thing or goods would be worth in case the peril was realised, but at how much the doubtful event should likely (*verisimiliter*) be estimated.
>
> (Santerna 1552/1971: p. 336; Franklin 2001: p. 277)

The estimation of prices was done by intuitive estimation of all the uncertain factors relevant to the individual case, rather than by collecting statistics. Shakespeare notes the rough quantification of risks by merchants, using the language of odds:

> Or what hath this bold enterprise brought forth
> More than that being which was like to be?
> We all that are engaged to this loss
> Knew that we ventured on such dangerous seas
> That if we wrought out life 'twas ten to one;
> And yet we ventured, for the gain proposed
> Choked the respect of likely peril fear'd.
>
> (Henry IV part II, 1.1.178–84)

Shakespeare uses "It's lots to blanks" to mean a near-certainty (Franklin 2001: pp. 280–3; Bellhouse and Franklin 1997). Various kinds of bets, speculations, and lotteries were common in business and wider circles, but the contrast between avidity of speculation and lack of theory is remarkable (e.g. Welch 2008).

The moral theologians of the early seventeenth century collected all such contracts under the classification "aleatory contracts", and wrote with clarity on the reality of the probabilistic entities that were being bought and sold. Thus Juan de Lugo (mentioned above in connection with probabilism) wrote, in a chapter on "gaming, wagers and insurance":

> The first condition for the justice of an insurance is, that the price be equal to the peril undertaken; certainly that the price paid for the obligation should be as much as that obligation is worth in the judgement of experts. This price is not a definite amount, but has a maximum, mean and minimum, as with buying and selling. As varied circumstances affect the peril, so the just price should be varied. The equality is to be taken from the quantity of the peril at the time of the contract, not after the event.
>
> (Juan de Lugo 1652: 2: p. 447; Franklin 2001: p. 288)

It is because probability was studied in this moral context that Pascal and Fermat posed questions about dice in a way that is strange to modern ears. The question they ask is about ethics: what is the just division of the stake in an interrupted game of chance? (Coumet 1970; Sylla 2003)

2.8 Dice

There is much archaeological evidence for ancient and medieval games of chance, but very little theory arose from this. There were occasional comments showing some basic understanding that some outcomes are more frequent than others, such as Aristotle's remark that "To succeed in many things, or many times, is difficult; for instance, to repeat the Koan throw ten thousand times would be impossible, whereas to make it once or twice is comparatively easy," but they were never developed into a quantitative theory.

A remarkable anonymous Italian manuscript of about 1400 poses the problem, very similar to that debated by Pascal and Fermat, of how the stake should be divided in an interrupted game, if one player has won two points and the other none, and the winner is the first to three. It is not an easy problem, as Pascal and Fermat discovered. The author solves it correctly with some complex reasoning from symmetries (and no explicit mention

of probabilities or counting of outcomes) (Toti Rigatelli 1985; Franklin 2001: pp. 294–6; background in Meusnier 2007a).

Gerolamo Cardano in the sixteenth century attacked such problems with some generality, and with an understanding of the idea that it involved counting the number of possible outcomes. However his work was often unclear, contained mistakes and was mostly unpublished until the late seventeenth century (Bellhouse 2005). Galileo correctly solved an individual problem involving the throws of three dice, but again his work was not published until centuries later (Franklin 2001: p. 302).

When Pascal and Fermat came to the problem, some knowledge of probability calculations was already current, but it is hard to discover what it was. The Chevalier de Méré, an amateur mathematician, posed to Pascal this conundrum: if I undertake to throw a 6 in four throws of a die, I have the advantage, with odds in my favour of 671 to 625. But if I undertake to throw a double 6 in 24 throws of two dice, I have the disadvantage. How is that possible? Pascal implies that de Méré was able to work this out for himself, which is not easy. The probabilities were plainly being calculated rather than gained from experience – indeed, no writer of the time suggests any connection between the odds being calculated and long-run relative frequency of throws. No method of working out the probabilities is simple.

Pascal and Fermat apply their well-developed mathematical skills to this problem and to the general problem of determining the just division of the stake in an interrupted game, for any position of the players. They succeed admirably. They say very little, however, about the nature of the entities they are calculating. In their initial letters on the just division of stakes, they merely calculate what would be "impartial" between the players. They appear to have no way of conceptualizing a probability except as a just share of a stake, a concept just sufficient for them to deploy the symmetry arguments that result in a numerical solution to the problem. Pascal prefers a mathematical argument involving recursion whereas Fermat prefers enumerating cases, but the difference is purely mathematical. In later letters, Pascal refers to two conditions as "equal and indifferent", while Fermat considers some imaginary cases so as to "make all the chances equal (*rendre tous les hasards égaux*)". That is all that is said. The words "probable" and "probability" do not appear (Franklin 2001: pp. 309–12).

Fermat does not think the topic is very interesting, being merely a source of not especially exciting problems in number theory. Only Pascal realizes that something remarkable has been discovered: precisely-calculable necessities in the variable material of chance: " . . . by thus uniting the demonstrations of mathematics to the uncertainty of chance, and reconciling what seem contraries, it can take its name from both sides, and rightly claim the astonishing title: the Geometry of chance (*aleae Geometria*)." What is calculated, however, is only "what is rightly due" to the players (Pascal 1964–70: 2: pp. 1034–5; Franklin 2001: p. 312).

2.9 Philosophical Lessons from the Pre-history of Probability

The story called the "pre-history" of probability has some lessons for the philosophy of probability. They are much the same lessons as arise from a developmental psychology

perspective, or a legal perspective, on probability. Reasoning under uncertainty is a very diverse field, comprising everything from subsymbolic risk evaluations in the rat brain to hunches on the demeanour of witnesses to the study of stochastic processes with advanced measure theory. A mathematical model of probability based on counting equiprobable outcomes of dice may or may not be applicable to some or all of the field. That is for debate in individual cases. The "pre-history" of probability, before that model was imposed, is a vast storehouse of examples of the various kinds of non-deductive arguments possible; the arguments are often in pristine form as suggested by real examples, before being deformed by inapplicable theory of one kind or another.

It is particularly significant that the law of evidence, which contributed more than any other discipline to the understanding of uncertain reasoning before Pascal, has almost entirely refused to accept quantification. Apart from some marginal and heavily qualified use of statistics in, for example, DNA identification cases, modern law has strongly resisted any attempts to quantify its probabilistic concepts such as "proof beyond reasonable doubt", and has refused all attempts to apply Bayesian formulas in court (Tribe 1971; Franklin 2011).

So early writers should not be seen as dealing in confused "anticipations" of the later theory of mathematical probability (or hardly ever). They are dealing with matters that in general are as unquantified now as they ever were – the degree to which evidence supports theory, the strength and justification of inductive inferences, the weight of testimony, the combination of pieces of uncertain evidence, the price of risk, the philosophical nature of chance, and the problem of acting in case of doubt (unquantified except in those cases where both probability and payoffs are well quantified). Those problems are even now the staple of philosophical discussion of probability.

References

Aquinas, Thomas (1952) *Disputed Questions on Truth.* Translated from the Latin by R. W. Mulligan. Chicago, IL: Regnery.

Aristotle (4th C BCE) *On the Heavens.* Translated from the Ancient Greek.

Aristotle (4th C BCE) *Rhetoric*. Translated from the Ancient Greek.

Azo of Bologna (1966) *Lectura super codicem*. Turin: Officina Erasmiana.

Bellhouse, D. (2005) Decoding Cardano's *Liber de Ludo Aleae*. *Historia Mathematica*. 32. 2. pp. 180–202.

Bellhouse, D. and Franklin, J. (1997) The language of chance. *International Statistical Review*. 65. 1. pp. 73–85.

Bonini, Alexander (1962) *Le tractatus de usuris de maître Alexandre d'Alexandria: un traité de morale économique au XIVe siècle.* Edited by A.-M. Hamelin. Louvain: Nauwelaerts.

Cano, M. (1734) *De locis theologicis.* In *Opere*. Padua: Typis Seminarii.

Ceccarelli, G. (1999) Le jeu comme contrat et le risicum chez Olivi. In Boureau, A. and Piron, S. (eds.) *Pierre de Jean Olivi (1248–1298): Pensée scolastique, dissidence spirituelle et société*. pp. 239–50. Paris: Vrin.

Ceccarelli, G. (2001) Risky business: theological and canonical thought on insurance from the thirteenth to the seventeenth century. *Journal of Medieval and Early Modern Studies*. 31. 3. pp. 607–58.

Ceccarelli, G. (2007) The price for risk-taking: marine insurance and probability calculus in the Late Middle Ages. *Journal Electronique d'Histoire des Probabilités et de Statistique*. 3. 1.

Cicero (c. 85 BCE) *De Inventione*.
Corpus Iuris Canonici (1879–81) Edited by A. Friedberg. Leipzig: Tauchnitz.
Corpus Iuris Civilis, Digest (1998) Translated in Watson, A. *The Digest of Justinian*. Philadelphia: University of Pennsylvania Press.
Coumet, E. (1970), La théorie du hasard est-elle née par hasard? *Annales: Économies, Sociétés, Civilisations.* 25. 3. pp. 574–98.
De Blic, J. (1930) Barthélémy de Medina et les origines du probabilisme. *Ephemerides Theologicae Lovanienses*. 7. pp. 46–83, 264–91.
de Lugo, Juan (1652) *De Iustitia et Iure*. Lyons: Borde, Arnaud, & Rigaud.
de Lugo, Juan (1868) *Disputationes scholasticae et morales.* Edited by J.-B. Fournials. Paris: L. Vivès.
Deman, T. (1936) Probabilisme. In Vacant, A., Mangenot, E., and Amann, E. (eds.) *Dictionnaire de Théologie Catholique*. XIII. 1. cols. 417–619. Paris: Letouzey et Ané.
Duns Scotus, J. (1304/1962) *Opus Oxoniense*. Translated in Wolter, A. *Duns Scotus: Philosophical Writings*. Edinburgh: Nelson.
Earman, J. (1992) *Bayes or Bust? A Critical Examination of Bayesian Confirmation Theory*. Cambridge, MA: MIT Press.
Fleming, J. (2006) *Defending Probabilism: The Moral Theology of Juan Caramuel*. Washington, DC: Georgetown University Press.
Franklin, J. (2001) *The Science of Conjecture: Evidence and Probability Before Pascal*. Baltimore, MD: Johns Hopkins University Press.
Franklin, J. (2009) *What Science Knows: And How It Knows It*. New York, NY: Encounter Books.
Franklin, J. (2011) The objective Bayesian conceptualisation of proof and reference class problems. *Sydney Law Review*. 33. 3. pp. 545–61.
Franklin, J. (2013) Probable opinion. In Anstey, P. R. (ed.) *The Oxford Handbook of British Philosophy in the Seventeenth Century*. Oxford: Oxford University Press.
Gabbay, D. M. and Koppel, M. (2011) Uncertainty rules in Talmudic reasoning. *History and Philosophy of Logic*. 32 .1. pp. 63–9.
Galen (1965) *Three Treatises on the Nature of Science.* Edited by M. Frede. Indianapolis, IN: Hackett.
Galileo (1967) *Dialogue Concerning the Two Chief World Systems.* Translated from the Italian by S. Drake. 2nd ed. Berkeley, CA: University of California Press.
Hacking, I. (2006) *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. 2nd ed. New York, NY: Cambridge University Press.
Hájek, A. (2002/2011) Interpretations of probability. Revised in Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/probability-interpret/ [Accessed 24 Aug 2015.]
Hobbes, T. (1839). *Human Nature*. In Molesworth, W. (ed.) *The English Works of Thomas Hobbes*. Vol. 4. pp. 1–76. London: John Bohn.
Jaynes, E. T. (2003) *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
Jonsen, A. R. and Toulmin, S. E. (1988) *The Abuse of Casuistry: A History of Moral Reasoning*. Berkeley, CA: University of California Press.
Kantola, I. (1994) *Probability and Moral Uncertainty in Late Medieval and Early Modern Times*. Helsinki: Luther-Agricola-Society.
Keynes, J. M. (1921) *A Treatise on Probability*. London: Macmillan.

Knebel, S. (1992) Necessitas moralis ad optimum (III): Naturgesetz und Induktionsproblem in der Jesuitenscholastik während des zweiten Drittels des 17. Jahrhunderts. *Studia Leibnitiana*. 24. 2. pp. 182–215.

Macdonald, R. R. (2004) Statistical inference and Aristotle's *Rhetoric*. *British Journal of Mathematical and Statistical Psychology*. 57. 2. pp. 193–203.

Maryks, R. A. (2008) *Saint Cicero and the Jesuits: The Influence of the Liberal Arts on the Adoption of Moral Probabilism*. Aldershot: Ashgate.

Mellor, D. H. (2005) *Probability: A Philosophical Introduction*. London: Routledge,.

Meusnier, N. (2007a) Le problème des partis peut-il être d'origine arabo-musulmane? *Journal Electronique d'Histoire des Probabilités et de Statistique*. 3. 1.

Meusnier, N. (2007b). À propos d'une controverse au sujet de l'interprétation d'un théorème "probabiliste" de Nicole Oresme. *Journal Electronique d'Histoire des Probabilités et de Statistique*. 3. 1.

Ohrenstein, R. and Gordon, B. (1991) Risk, uncertainty and expectation in Talmudic literature. *International Journal of Social Economics*. 18. 11. pp. 4–15.

Olivi, P. J. (1980) *Un trattato di economia politica francescana: il De emptionibus et venditionibus, de usuris, de restitutionibus di Pietro di Giovanni Olivi*. Rome: Istituto storico italiano per il Medio Evo.

Oresme, N. (1966) *De proportionibus proportionum* and *Ad pauca respicientes.* Translated from the Latin by E. Grant. Madison, WI: University of Wisconsin Press.

Pascal, B. (1964–70) *Oeuvres Complètes*. Bruges: De Brouwer.

Piron, S. (2007) Le traitement de l'incertitude commerciale dans la scolastique médiévale. *Journal Electronique d'Histoire des Probabilités et de Statistique*. 3. 1.

Prierias, Sylvester (1515) *Summa summarum*. Bologna.

Ptolemy (1984) *Almagest*. Translated from the Ancient Greek by G. J. Toomer. New York, NY: Springer.

Rabinovitch, N. L. (1973) *Probability and Statistical Inference in Ancient and Medieval Jewish Literature*. Toronto: Toronto University Press.

Santerna, P. (Pedro de Santarém) (1552/1971) *De assecurationibus et sponsionibus mercatorum*. Edited by M. Amzalak. Lisbon: Greìmio dos Seguradores.

Schüssler, R. (2005) On the anatomy of probabilism. In Kraye, J. and Saarinen, R. (eds.) *Moral Philosophy on the Threshold of Modernity*. pp. 91–113. Dordrecht: Springer.

Schwartz, D. (2014) Probabilism reconsidered: deference to experts, types of uncertainty, and medicines. *Journal of the History of Ideas*. 75. 3. pp. 373–93.

Sextus Empiricus (1996) *Outlines of Pyrrhonism*. In *The Skeptic Way*. Translated from the Latin by B. Mates. New York, NY: Oxford University Press.

Suárez, F. (1944) *Selections from Three Works of Francisco Suárez, S.J.* Translated from the Spanish by G. L. Williams et al. Oxford: Clarendon.

Sylla, E. D. (2003) Business ethics, commercial mathematics, and the origins of mathematical probability. *History of Political Economy*. 35. Annual Supplement. pp. 309–37.

Toti Rigatelli, L. (1985) Il "problema delle parti" in manoscritti del XIV e XV secolo. In Folkerts, M. and Lindgren, U. (eds.) *Mathemata: Festschrift für Helmut Gericke*. Stuttgart: Steiner.

Tribe, E.L. (1971) Trial by mathematics: precision and ritual in the legal process. *Harvard Law Review*. 84. 6. pp. 1329–93.

Welch, E. (2008) Lotteries in early modern Italy. *Past and Present*. 199. 1. pp. 71–111.

CHAPTER 3

PROBABILITY IN 17TH- AND 18TH-CENTURY CONTINENTAL EUROPE FROM THE PERSPECTIVE OF JACOB BERNOULLI'S *ART OF CONJECTURING*

EDITH DUDLEY SYLLA

About the year 1689, Jacob Bernoulli, professor of mathematics at Basel, entered into his research journal the proof of an important theorem that was to be the last item in his *Ars Conjectandi* (*The Art of Conjecturing*), a book not published – and then in incomplete form – until 1713, eight years after his death in 1705. After the proof, he wrote:

> Nota Bene. I esteem this discovery more than if I had given the quadrature of the circle itself, which even if it were found very great, would be of little use.
>
> (*N.B. Hoc inventum pluris facio quam si ipsam circuli quadraturam dedissem, quod si maximè reperiretur, exigui usûs esset*.
>
> (Bernoulli 1975: p. 88)

Siméon-Denis Poisson later gave Bernoulli's theorem the name 'the law of large numbers' (now qualified as 'the weak law of large numbers'), but in this chapter, where there is no danger of confusing it with Daniel Bernoulli's theorem in fluid mechanics, I will call it simply 'Bernoulli's fundamental theorem' or 'Bernoulli's theorem', since I think that Bernoulli would not have considered his theorem to be a law.

As it was published in 1713, *Ars Conjectandi* had four parts. Part I reproduced Christiaan Huygens' *De Ratiociniis in Ludo Aleae* (1657), along with Bernoulli's extensive notes. Part II covered combinations and permutations and Part III the application of combinations and permutations to games of chance. Part IV applied the mathematics of the preceding

parts by analogy to civil, moral, and economic matters. In the fifth and last chapter of Part IV, Bernoulli demonstrated his fundamental theorem showing how ratios of cases for and against a given outcome might be learned a posteriori or from experience.

In his choice of the title *Ars Conjectandi*, Bernoulli signaled that what he hoped to produce was not a science, but a practical art, one which might enable a person to make wise decisions, while admitting that certainty is impossible. Bernoulli defined the art of conjecturing in terms of probabilities:

> To *conjecture* about something is to measure its probability. Therefore we define the *art of conjecture*, or *stochastics*, as the art of measuring the probabilities of things as exactly as possible, to the end that, in our judgments and actions, we may always choose or follow that which has been found to be better, more satisfactory, safer, or more carefully considered. On this alone turns all the wisdom of the philosopher and all the practical judgment of the statesman.
>
> Probabilities are assessed according to the *number* together with the *weight* of the *arguments* that in any way prove or indicate that something is, will be, or has been. By *weight* I mean probative force.
>
> (Bernoulli 2006: pp. 317–18; Bernoulli 1713: pp. 213–14)

Mathematics was to be the preeminent tool of the art of conjecture because it enabled one to reason infallibly from premisses or principles to conclusions. But this did not require that the premisses or principles in themselves be certain, and in fact it was assumed that in this case they were as a rule not certain. According to Aristotle and to many late medieval and early modern Aristotelians, scientific knowledge is knowledge of conclusions demonstrated on the basis of true principles. As in Euclidean geometry, these principles might be axioms (principles accepted widely by rational people as self-evident) or they might be postulates (principles accepted as certain by practitioners of the given discipline, but not self-evident). In disciplines such as astronomy and optics, many of the basic principles were based on sense and experience. In disciplines such as astrology, the basic principles might be highly uncertain and, as a result, the conclusions also uncertain. (See Bernoulli 1744/1967: pp. 233–4.)

In accordance with this picture, disciplines such as ethics do not demonstrate the truth of their conclusions with certainty, but at best make probable arguments. At the beginning of his *Nicomachean Ethics*, Aristotle wrote:

> We must be content, then, in speaking....about things which are only for the most part true and with premisses of the same kind to reach conclusions that are no better.
>
> (Aristotle, Nicomachean Ethics 1941: 1094b19–23; cf. Sylla 1990).

In scholastic Latin *probabilis* was a term applied to propositions and arguments that were considered plausible or rationally defensible, but not certain or demonstrable. As it was related to conjecture, Bernoulli likewise understood probability to refer to less than complete certainty:

> *Probability*, indeed, is degree of certainty, and differs from the latter as a part differs from the whole. Truly, if complete and absolute certainty, which we represent by the letter *a* or by 1, is supposed, for the sake of argument, to be composed of five parts or probabilities, of which

> three argue for the existence or future existence of some outcome and the others argue against it, then that outcome will be said to have $3a$ / 5 or 3/5 of certainty. One thing therefore is called *more probable* than another if it has a larger part of certainty, even though in ordinary speech a thing is called probable only if its probability notably exceeds one-half of certainty
>
> (Bernoulli 2006: pp. 315–16).

This definition, it should be noted, appeared only in Part IV of *Ars Conjectandi*.

In Parts I–III of the book, on the other hand, Bernoulli had developed a set of mathematical tools that were applied to calculating expectations in games of chance. These tools ranged from simple algebra to infinite series and the building up of facility in dealing with combinations and permutations. The mathematics in Parts I–III was not new, but was an extension of mathematical tools already included in commercial arithmetic or reckoning books, supplemented in places by classical Greek mathematics. Thus the main formula for calculating expectation in a game of chance where the player might receive one or another prize under various conditions had the same form as the rule for the unit value of a mixture – the so-called rule of mixture or *regula alligationis* (Bernoulli 2006: pp. 57, 138). In theses he had proposed for disputation in competition for the open chair in mathematics at Basel in 1687, Bernoulli had noted that algebra can be used to provide as many new rules as one might wish (Bernoulli 1744/1967: p. 296).

In *Ars Conjectandi* as it was eventually published, Bernoulli had made relatively little progress in Part IV in showing how the art of conjecturing could work in civil, moral, and economic matters before he broke away from his previous line of thought to state and prove his fundamental theorem. Nevertheless, it was *Ars Conjectandi*, or the reports of it that existed before its publication in 1713, that laid the foundation for mathematical probability so named, as distinct from the mathematics of games of chance.

3.1 The Perspective from Bernoulli's *Ars Conjectandi*

As a discipline, probability is now part of mathematics. It was first axiomatized by A. N. Kolmogorov in 1933. Like Euclidean geometry, mathematical probability can prove theorems that follow from its definitions, axioms, and postulates, perhaps extended by newly devised approaches and methods consistent with its axioms. In the late seventeenth and early eighteenth centuries the discipline of mathematical probability began to be assembled from pieces of mathematics that previously had had little or nothing to do with each other. One piece was the mathematics of expectation in games of chance, represented most prominently by Christiaan Huygens' *De ratiociniis in ludo aleae* (1657; *On Reckoning in Games of Chance*). Another part was the mathematics of combinations and permutations. Resources also came from algebra, the mathematics of infinite series, the use of logarithms, the drawing of curves, and so forth. Some of these resources were parts of abstract mathematics, such as arithmetic or geometry, but many of them had previously been found in works of practical, applied, or concrete mathematics, specifically in books of commercial arithmetic.

According to B. L. van der Waerden (1975: p. ix), Christiaan Huygens and Jacob Bernoulli created the theory of probability nearly *ex nihilo* 'fast aus dem Nichts geschaffen'). Girolamo Cardano, Pierre Fermat, and Blaise Pascal had calculated probabilities and expectations for a few games of chance, van der Waerden wrote, but no one before Huygens had tried to present an organized theory.

Ironically, the influence of *Ars Conjectandi* began before its publication. In 1709, after Jacob Bernoulli's death, but before the publication of *Ars Conjectandi*, Jacob's nephew Nicolaus I Bernoulli (given the numeral I by historians to distinguish him from other mathematicians with the same name), in his master's thesis for a degree in law entitled *De usu artis conjectandi in jure* (*On the use of the art of conjecturing in law*), had examined some of the sorts of problems that Jacob might have had in mind for the rest of Part IV, for instance estimating life expectancy, determining when someone absent for a long time might be declared dead, valuing inheritances, pricing insurance, and evaluating the credibility of witnesses. Nicolaus I published a report of his law dissertation in the 1711 *Acta Eruditorum*.

At Jacob's death, his student, Jacob Hermann, had been asked to go through his papers and to provide information to those who intended to prepare eulogies. In his *Essay d'analyse sur les jeux de hazard* (1708; *Essay of analysis on games of chance*), Pierre Rémond de Montmort transmitted the information from the eulogies. Mathematicians during the past fifty years had made great achievements in applying mathematics to physics, he said, but it would be even more glorious if mathematics could serve to rule judgments and conduct in practical life. This is what Bernoulli had proposed to do in his book, to be named *De arte conjectandi, l'art de deviner*, if his premature death had not prevented it. Fontenelle and Saurin had each given a short description of the proposed book in their *éloges* published in the *Histoire de l'Academie* (1705) and in the *Journaux des Sçavans de France* (1706) respectively. Although he did not know for which games Bernoulli had determined the relative shares (*partis*) of the players, Montmort proposed to help fill the gap left because of Bernoulli's early death by the publication of his own work. He himself, however, found it too difficult to deal with practical decision-making, and so concentrated on games and on the mathematics of combinations and permutations. Montmort's substantial book on games was first published, anonymously, in 1708.

Not long thereafter, Abraham De Moivre published in Latin, in the *Philosophical Transactions of the Royal Society of London* for 1712, a shorter work with the title *De mensura sortis, seu, de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus* (*On the measure of lot, or, of the probability of outcomes in games depending on chance*). In a letter to Francis Robartes prefaced to *De mensura sortis*, De Moivre gave this faint praise to Montmort's *Essay d'analyse*:

> Huygens was the first that I know who presented rules for the solution of this sort of problems, which a French author has very recently well illustrated with various examples; but these distinguished gentlemen do not seem to have employed that simplicity and generality which the nature of the matter demands; moreover, while they take up many unknown quantities to represent the various conditions of gamesters, they make their calculations too complex; and while they suppose that the skill of the gamesters is always equal, they confine this doctrine of games within limits too narrow.
>
> (Hald 1990: p. 291; Hald is translating from Montmort's quotation of this passage)

In what follows, then, I will concentrate on Bernoulli's *Ars Conjectandi* together with Christiaan Huygens' *De ratiociniis in ludo aleae*. At the end of the chapter, I will refer to Abraham De Moivre's *De Mensura Sortis* (1712) and to his *The Doctrine of Chances* (1718; 2nd edition 1738; 3rd edition 1756) because only in De Moivre's work is the definition of probability in a frequentist sense fully established. I focus on Bernoulli's work in order to make clear both the mathematics and the conceptual context of this first major work of mathematical probability. I set aside the writings of Pascal, Fermat, and Montmort on games of chance in order to discuss Jacob Bernoulli's *Ars Conjectandi* in sufficient detail to make its special characteristics clear. For a complementary statement of my views on the role of Jacob Bernoulli's *Ars Conjectandi* in the emergence of mathematical probability, see Sylla 2014.

3.2 The Publication of *Ars Conjectandi*

Jacob Bernoulli's *Ars Conjectandi* (1713/1968) closely reproduced the manuscript he had left behind at his death in 1705. His wife and *son* Nicolaus, a painter, kept control of the manuscript from the time of Jacob's death until it was turned over to the printer. When Jacob's *nephew* Nicolaus I returned to Basel after his travels, he found that the printing was nearly complete. At Pierre Varignon's suggestion, Nicolaus I proofread the printing but he found only a few errata, because, as he said, the manuscript had been very neat and well written (*le manuscrit ayant été fort net et tres bien écrit*). I mention this to correct those who imagine that Nicolaus I had an important role in putting together *Ars Conjectandi*. Although the publishing history of *Ars Conjectandi* has been thoroughly documented by K. Kohli (Bernoulli 1975: pp. 391–401) and is reported in Yushkevich 1987, historians repeatedly, but erroneously, credit Nicolaus I Bernoulli with editing the work. Let us now examine *Ars Conjectandi* in more detail.

3.3 Part I of *Ars Conjectandi*

In Part I of *Ars Conjectandi* Bernoulli explains and extends Huygens' *De ratiociniis in ludo aleae*. If the algebraic formulas found in Huygens' *De ratiociniis* and, consequently, in Part I of *Ars Conjectandi* were similar to rules that had been found in works of commercial arithmetic, the fundamental concept used in applying these formulas to games of chance was 'lot' (*sors*) or expectation (*expectatio*). Huygens had opened *De ratiociniis* by saying:

> Although the outcomes of games that are governed purely by lot (*sors*) are uncertain, the extent to which a person is closer to winning than to losing always has a determination. Thus, if a person undertakes to get a six on the first toss of a die, it is indeed uncertain whether he will succeed, but how much more likely he is to fail than to succeed is definite and can be calculated. Similarly, if I were to contend with someone on the understanding that three games are needed to win, and I had already won one game, it would still be uncertain which of us would win three games first. Yet we can calculate with the greatest certainty how great

> my expectation (*expectatio*) and my opponent's expectation should be appraised to be. From this we can also determine how much greater a share (*portio*) of the stakes I should get than my opponent if we agree to quit with the game unfinished, or how much should be paid by someone who wanted to continue the game in my place and with my lot.
>
> (Bernoulli 2006: p. 132; Bernoulli 1713/1968: p. 3)

He then states the fundamental principle that lies behind his calculations as follows:

> I use the fundamental principle that a person's lot (*sors*) or expectation to obtain something in a game of chance should be judged to be worth as much as an amount such that, if he had it, he could arrive again at a like lot or expectation contending under fair conditions.
>
> (Bernoulli 2006: p. 133; Bernoulli 1713/1968: pp. 3–4)

In his note to this passage, Bernoulli tries to explain Huygens' principle in more popular terms:

> I will try to demonstrate it by reasoning that is more popular than the previous and more adapted to common comprehension. I posit only this as an axiom or definition: *Anyone may expect, or should be said to expect, just as much as he will acquire without fail.* . . . [*unusquisque tantundem expectet, vel expectare dicendus sit, quantum infallibiliter obtinebit*]. It can be seen from what we have said that we are not using the word *expectation* in its ordinary sense, according to which we are commonly said to expect or to hope for what is best of all, though worse things can happen to us. Here account is taken of the extent to which our hope of getting the best is tempered and diminished by fear of getting something worse. So by its 'value' we always mean something intermediate between the best we hope for and the worst we fear.
>
> (Bernoulli 2006: p. 134; Bernoulli 1713/1968: p. 5)

Before Huygens, the key Latin word in analyzing games of chance was *sors* or 'lot'. Because games, like business partnerships, were supposed to be fair, what one paid to participate in a game of chance was supposed to equal the value of what one had thereby purchased. It was not an accident that the word *sors* was also used to refer to the capital of an investment. In translating Huygens' work from the Dutch, in which it was originally written, into Latin, Frans van Schooten introduced *expectatio* as an alternative or substitute for *sors*, first writing '*sortem seu expectationem*', and then simply '*expectationem*', when in the original Dutch Huygens had written *kans* or *kansse*, 'chance' or 'chances' or words meaning 'value' or 'worth'. Thus van Schooten redefined '*expectatio*' in his translation of Huygens' work, as Bernoulli pointed out in his note, and at the same time replaced '*sors*', the word that had previously been used in this context, with '*expectatio*'.

Before Huygens, Girolamo Cardano, Blaise Pascal, and others had also used *sors* as the key concept in relation to games of chance. It is now recognized by historians that before the flurry of works related to mathematical probability that appeared in the first decades of the eighteenth century, the key concept in the mathematics of games was expectation and not probability (Daston 1988). This was true already before Huygens, but the Latin word in use was '*sors*' or 'lot', rather than '*expectatio*'. The grounds for determining lot or expectation was not probability in the sense of frequency, as in later authors, but rather equity or fairness (cf. Bellhouse 2005). If two players are tossing a fair coin and the one betting on heads should get the whole money paid in if heads appears, and similarly for the one betting on tails, then

each player should put in an equal amount, and each player's lot or expectation will be equal to half of the total put in, in other words just what he put in. If x is what each player pays to play, then each player's expectation equals $\frac{1}{2}(2x) + \frac{1}{2}(0) = x$.

This way of thinking about games of chance stems from the earlier mathematics of partnerships or companies as set out in numerous books of commercial arithmetic. In a fair business contract, the investment, in relation to what other partners invested – in money, labor, or equipment (and in relation to the time periods involved, if this differed among the investors) – was considered in determining the investor's fair share of the profits, the terms of which would be set down in a notarized contract establishing the partnership or company and determining the distribution of profit or loss at the end of a pre-established time period.

European commercial arithmetic books often contained story problems that were inherited from Islamic algebra books. Among such problems were problems of how to divide the capital and profits of a company or partnership if the company was dissolved before the contracted time period (see Sylla 2003, 2006b). In Islam, these problems arose because the heirs of a deceased partner were entitled to ask for their respective inheritances. According to Islamic law, various relatives were entitled to certain proportions of the estate of the deceased, so that in Islam commercial law intersected with inheritance law. When the relevant mathematical works were translated from Arabic to Latin or vernacular languages, the concern for fairness to heirs generally meant that the partners' shares in the company that was being dissolved prematurely were evaluated taking into account how much of the duration initially contracted had been completed. Reasoning along these lines is carried over into the *Summa de arithmetica geometria proportioni & proportionalita* (1494) of Luca Paccioli and applied to games.

In Europe, however, this reasoning might be questioned. In his *Practica arithmetice*, published in 1539, Girolamo Cardano argued that Luca Paccioli erred in solving the division problem for a game of chance by looking to the past:

> [Luca Paccioli] errs in the determination of games with a very manifest error, even recognizable by a child . . . . Thus in playing a game to 6, where one player has 5 and the other 2, after many superfluous calculations, he says that the parts are 5 and 2, such that he divides the whole sum in 7 parts. Let us suppose that two players play to 19 and one has 18 and the other 9. Therefore [according to Paccioli] the first receives 2/3 of the whole sum and the other 1/3. Let the deposit of each be 12 gold pieces, with a total of 24. Then the first receives 16 and the second 8. This means that the player who has won 18 games has received from the second player only 4 gold pieces, which is a third of his deposit. But he needs only one more game to win, while the second player needs 10 games. This is most absurd. Moreover, [in the division] each player should take that part which on a fair basis (*aequa ratione*) he could deposit for that condition. But someone who has 18, playing with someone who has 9, when the limit is 19, could deposit 10 or even 20 against 1. Therefore in the division he should have 20 parts and the other player only 1 part.
>
> (Cardano 1539/1662: p. 214).

Here Cardano does not try to make an exact calculation. In an earlier passage, Cardano had made a calculation using the numbers of games each player lacks, although again he did not get the correct answer by modern standards. In looking to the future, Cardano took the same approach as would be taken by Pascal, Fermat, and Huygens, although he did not get the correct answer in later terms.

Frequently Cardano's contribution to the development of mathematical probability is put into doubt and, instead, correspondence between Blaise Pascal and Pierre de Fermat that took place in 1654 is taken to have been the fountainhead of mathematical probability (see Edwards 2002, 2003; Meusnier and Piron 2007). Pascal and Fermat are given credit as founders because they looked to the future of the game rather than the past, and, perhaps more important, because they got the right answer. In this chapter, in order to conserve space, I do not discuss the correspondence of Pascal and Fermat in detail. What was in their work was also, independently, in Huygens' *De ratiociniis*, and Huygens' work was the source used by Bernoulli (Huygens knew about the correspondence of Pascal and Fermat, and he knew their solutions to the relevant problems, but he did not know their methods; cf. Huygens' prefatory letter to van Schooten in *De ratiociniis*, as in Bernoulli 2006: p. 132). The origin of mathematical probability involved a conversation among multiple participants and by no means a unique transmission via a few great contributors.

To return to my discussion of Part I: in his note on Huygens' Proposition IV, Bernoulli repeats a point like that made by Cardano, concerning looking only to the future:

> A. [Huygens:] *For the number of games that each side lacks must be considered.* [Bernoulli:] In general we should take no account of past games when we compute the lots for games that are all in the future. For in any new game the probability that fortune will continue to favor those that it has favored before is no greater than the probability that it will favor those who have been the most unfortunate.
>
> (Bernoulli 2006: p. 139; Bernoulli 1713/1968: p. 12)

This passage includes two of Bernoulli's rare uses of 'probability' in Part I. It should be noted that Bernoulli uses 'probability' when he is writing about fortune and not about a stable underlying frequency of occurrence.

Thus, as far as Part I of *Ars Conjectandi* is concerned, Bernoulli is working within the framework of Huygens' *De ratiociniis*, and his approach is shaped by ideas of fairness and expectation, ideas that are shared by earlier writers on partnerships, companies, and games. In fitting the details of games to the available rules or equations of commercial arithmetic (or in making up new rules using algebra), mathematicians had to fashion concepts that could be related to the mathematical relationships proposed. Van Schooten's introduction of *expectatio* (originally nearly a synonym for *sors*, but gradually diverging in meaning), together with its link to a letter symbol in an equation, helped to promote the gradual distinction between the amount expected in a given case and the relative frequency with which similar cases occurred. Then 'probability' could evolve from meaning 'degree of certainty' to meaning 'relative frequency'.

As Bernoulli continues to comment on Huygens' text, he alternates between pointing out ways in which the problems may be framed to make the solution possible, on the one hand, and suggesting ways of organizing the calculations so that they are less prone to error. Besides Huygens' methods, he proposes additional methods to solve the same problems (Bernoulli 2006: pp. 157–8; Bernoulli 1713/1968: pp. 30–31). Frequently, he introduces symbols for generality. In the course of discussing Proposition XII, he finds a solution from the intersection of a straight line and a logarithmic curve (Bernoulli 2006: p. 168; Bernoulli 1713/1968: p. 43). Concerning Proposition XIV, Bernoulli notes:

> The Author [Huygens] in this Problem is first compelled to employ algebraic analysis, while in the preceding only synthesis was used. The difference between these two is that in all the

> former propositions the expectation sought was derived from other expectations that were either totally known and given, or, indeed not known, but naturally prior and simpler, and not dependent in turn upon that sought …. Here, however, the matter is different, for the expectation that I possess when it is my opponent's turn to throw cannot be estimated in the Author's customary way unless I take as known the lot that I acquire when it is my turn to throw. But I also cannot find out the latter unless I take the former to be already found out, which is the very same thing that I intend to discover ….
>
> (Bernoulli 2006: p. 172: Bernoulli 1713/1968: pp. 47–8)

And then he proposes another way to solve the problem that makes use of infinite series (Bernoulli 2006: p. 173: Bernoulli 1713/1968: pp. 48–9). What is important here is the clear and orderly presentation of mathematical methods, rather than any new conceptualization of probability or frequency.

3.4 *Ars Conjectandi*, Parts II and III

Part II of *Ars Conjectandi*, dealing with combinations and permutations, represents a significant part of Jacob Bernoulli's contribution to the development of mathematical probability. It was important insofar as it provided technical assistance in managing large numbers of alternative possibilities systematically. A real breakthrough in thinking about games broken off prematurely was to develop a way to enumerate systematically all the equally likely alternative outcomes and then to examine them to see which players would have won what in various scenarios. This is the approach used by Huygens and Bernoulli in Part III of *Ars Conjectandi*.

Developing the mathematics of combinations and permutations was essential if there was to be any progress in applying mathematics to the complexity of civil, moral, and economic matters. Already, in a note in Part I, Bernoulli wrote:

> Unless one investigates the number of throws in some orderly manner, it is easy to overlook one or more throws, especially if there are many dice. So I shall show what method we should use to be sure we find all the throws and overlook none.
>
> (Bernoulli 2006: p. 147; Bernoulli 1713/1968: p. 21)

> Since this method of calculating the number of throws with several dice is inordinately tedious and lengthy, I shall now present a method that accomplishes the same thing not only for a certain number of points, but for all numbers of points. This method uses the adjoined Table, which can be constructed quite easily and which makes readily visible the nature of these numbers and the relations among them. This is how the Table is constructed ….
>
> (Bernoulli 2006: pp. 149–50; Bernoulli 1713/1968: pp. 23–4)

3.5 *Ars Conjectandi*, Part IV

In Part IV, Bernoulli finally takes up the topic of probability, so named. The first four chapters of Part IV, applying the mathematics of Parts I–III to issues of probability taken in an

epistemic sense, are not an impressive achievement, largely because they take on problems that even today have not been successfully mathematized. If we assume that the available mathematics is what was displayed in Parts I–III, then Bernoulli's problem was how to fit that mathematics to civil, moral, and economic subject matter. In an exchange of letters with Leibniz, Bernoulli had suggested his method might apply to weather forecasting or to the calculation of life expectancies, where indeed there was a potential for his mathematical approach at some future time, although not in his lifetime, given the lack of relevant accumulated data and the difficulty of forming concepts to fit the data to equations (see Franklin 2001: p. 332 re Huygens and 'translation axioms' to relate the real world to the mathematics).

Other examples of potential applications are even less promising. There had long been a minor use of mathematics in law with regard to so-called 'indices'. According to typical reasoning, to prove a felony a confession or two eye-witnesses were required. Failing either of these, the accused might be tortured to elicit a confession, but torture was not authorized unless a 'half proof' could be provided. So, for example, if two eye-witnesses constitute a proof, then one eye-witness might count as a half proof. Bernoulli suggests combining diverse kinds of evidence in a mathematical formula to produce such a partial proof or probable conjecture. He writes:

> The *arguments* are either *internal* or *external.* Internal or, as they are more commonly called, technical arguments are taken from the topics – cause, effect, subject, associated circumstances, sign, or anything else that seems to have a connection with the thing to be proved. External and nontechnical arguments appeal to human authority and testimony. Here is an example. Titius is found slain on the road. Maevius is accused of having committed the murder. The arguments for the accusation are: 1. It is established that Maevius hated Titius. (This is an argument from *cause*, for hatred could have driven him to kill Titius.) 2. When Maevius was questioned, he turned pale and answered timidly. (This is an argument from *effect*, for his pallor and fear may have resulted from his consciousness of having committed the crime.) 3. A sword stained by blood was found in Maevius's house. (This is a *sign*) ....[and so forth]
>
> (Bernoulli 2006: p. 318; Bernoulli 1713/1968: p. 214)

Here the 'topics' that Bernoulli refers to are headings employed by rhetoricians in order to come up with arguments in a systematic way. But how could all of these disparate factors be combined mathematically to come up with a number representing the probability that Maevius is guilty of the murder? Bernoulli tries to sort through the evidence and classify it before trying to combine it. He distinguishes between arguments that exist necessarily and those that exist contingently, and then between those arguments that indicate necessarily and those that indicate contingently. He gives the example:

> according to the rules of a game, a dice player wins if he throws a seven with two dice. I want to conjecture how much hope he has of winning. Here the argument for his winning is a throw of seven, which indicates necessarily (the necessity established, to be sure, by the contract entered into by the players) but exists only contingently, since points other than seven may fall.
>
> (Bernoulli 2006: pp. 321–2; Bernoulli 1713/1968: p. 218)

Then Bernoulli distinguishes further between *pure* arguments, which are such that they prove a thing in some cases, but prove nothing positively in other cases, and *mixed*

arguments, which prove a thing in some cases in such a way that they prove the contrary in the other cases. After making these distinctions, Bernoulli proposes complicated sums and differences of fractions to represent the combined evidence, but finally admits:

> I cannot conceal here that I foresee many problems in particular applications of these rules that could cause frequent delusions unless one proceeds cautiously in discerning arguments. For sometimes arguments can seem distinct that in fact are one and the same. Or, vice versa, those that are distinct can seem to be one. Sometimes what is posited in one argument plainly overturns a contrary argument.
>
> (Bernoulli 2006: p. 325; Bernoulli 1713/1968: pp. 221–2).

Not only was there a problem of enumerating separate arguments, but there was a problem of measuring the weight of these arguments on any kind of homogenous scale. In these circumstances, it is not surprising that when Bernoulli came up with the idea of finding the ratios of cases a posteriori or by experience, he jumped at the possibility that by following the a posteriori approach he could leave behind these rather unpersuasive formulas:

> .... the only thing needed for correctly forming conjectures on any matter is to determine the numbers of these cases accurately and then to determine how much more easily some can happen than others. But here we come to a halt, for this can hardly ever be done. Indeed, it can hardly be done anywhere except in games of chance …. But what mortal, I ask, may determine, for example the number of diseases as if they were just as many cases, which may invade at any age the innumerable parts of the human body and which imply our death? And who can determine how much more easily one disease may kill than another....? Who, then, can form conjectures on the future state of life and death on this basis?
>
> (Bernoulli 2006: pp. 326–7; Bernoulli 1713/1968: pp. 223–4)

It is at this point that Bernoulli proved his fundamental theorem. Introducing the proof, he wrote:

> What cannot be ascertained a priori, may at least be found out a posteriori from the results many times observed in similar situations, since it should be presumed that something can happen or not happen in the future in as many cases as it was observed to happen or not to happen in similar circumstances in the past …. This empirical way of determining the number of cases by experiments is neither new nor uncommon.... everyone consistently does the same thing in daily practice. Neither should it escape anyone that to judge in this way concerning some future event it would not suffice to take one or another experiment, but a great abundance of experiments would be required …. But although this is naturally known to everyone, the demonstration by which it can be inferred from the principles of the art is hardly known at all, and, accordingly, it is incumbent upon us to expound it here. But I would consider that I had not achieved enough if I limited myself to demonstrating this one thing of which no one is ignorant. Something else remains to think about, which perhaps no one has considered up to this point. It remains, namely, to ask whether, as the number of observations increases, so the probability increases of obtaining the true ratio between the numbers of cases in which some event can happen and not happen, such that this probability may eventually exceed any given degree of certainty.
>
> (Bernoulli 2006: pp. 327–28; Bernoulli 1713: pp. 224–25)

The proof of an affirmative answer to this question constituted the fundamental theorem of which Bernoulli was so proud.

In the mathematics of games of chance, the different possible cases or outcomes are known by the make-up of the game pieces. A die may fall with any of the numbers from one to six facing up. And so forth. In nature, Bernoulli asserts, there are similar alternative possible cases or outcomes, although they are not apparent to us. We know this to be true, because whatever God knows is certain:

> In themselves and objectively, all things under the sun, which are, were, or will be, always have the highest certainty. This is evident concerning past and present things, since, by the very fact that they are or were, these things cannot not exist or not have existed. Nor should there be any doubt about future things, which, in like manner, even if not by the necessity of some inevitable fate, nevertheless by divine foreknowledge and predetermination, cannot not be in the future. Unless, indeed, whatever will be will occur with certainty, it is not apparent how the praise of the highest Creator's omniscience and omnipotence can prevail.
>
> (Bernoulli 2006: p. 315; Bernoulli 1713/1968: pp. 210–11)

If we can calculate expectations in games of chance on the basis of the alternative cases or possibilities (as in the numbers one to six on the faces of a die), then so too can we calculate with regard to practical civil, moral, or economic problems, if we make enough observations in similar circumstances to learn, at least approximately, the ratios of alternative possibilities. So we might calculate life expectancies for twenty-year-old men by observing what happened to 200 or 20,000 twenty-year-old men living in similar circumstances as they grew older until they eventually died.

Scholars have disagreed about the proper interpretation of Bernoulli's fundamental theorem (Hacking 1975: p. 154). For Bernoulli, I think that the theorem was, in effect, a proof of concept, as is clear from the language at the end of Bernoulli's introduction to his proof, quoted above. If, although hidden from us, there is some fixed ratio of possible outcomes underlying the phenomena, then that ratio will emerge from observation – not the exact ratio but within some interval – with as high a probability as we choose, provided that we make enough observations. In the particular example that Bernoulli chose in proving his theorem, the two possible underlying factors were supposed to be in a ratio of 3 so-called fertile cases to 2 so-called sterile cases. Then for it to be 1000 times more probable that the ratio of fertile outcomes observed to all the outcomes observed would be not more than 31 to 50 nor less than 29 to 50, Bernoulli calculated that it would be necessary to make at least 25,550 observations (Bernoulli 2006: p. 339; Bernoulli 1713/1968: p. 238).

Commentators have suggested that Bernoulli did not continue further with his book manuscript because he was discouraged by how large a number of observations or experiments would be necessary to attain reasonable certainty about the results (Stigler 1986: p. 77). Soon after Jacob Bernoulli, others proved similar theorems that required considerably fewer observations for the desired level of certainty (De Moivre proposed a new approach in 1733, saying that Jacob and Nicolaus Bernoulli had determined very wide limits. See Hald 1990: pp. 266–7, 485–95). I doubt that Bernoulli was discouraged by the number, 25,550, of observations needed, on the grounds that Bernoulli claimed to value his accomplishment so highly. Bernoulli himself said that he was slow to finish because of his natural laziness and poor health (Bernoulli 1993: p. 116). In addition, his correspondence with G. W. Leibniz in the last years of his life shows that Bernoulli was still searching for

good examples or data to use, for instance examples that he hoped to find in a pamphlet on annuities written in Dutch by Jan de Witt (Sylla 1998, 2014).

To prove his theorem, Bernoulli first proved five lemmas related to the series expansion of integral powers of binomials. Then he related the terms of the expansion to the 'degrees of probability or the numbers of cases' in which all the outcomes of an experiment are fecund, or all fecund except 1 or 2 or 3 or 4 etc. sterile ones (Bernoulli 2006: p. 338; Bernoulli 1713/1968: p. 338; see Sylla 1997). Bernoulli's achievement here was to see how he could construct a mathematical expression that could be related to the outcomes of trials of a process with *r* cases for one outcome and *s* cases for the opposite, for a total of *t* cases. He does not ask how likely it is that unknown processes would fit such a binomial model. The word 'probability', which had not or only rarely been relevant to Parts I–III, occurs in Part IV only in relation to the epistemic probability that the observed ratio of outcomes would fall within a small interval around the true ratio of cases after a certain number of observations.

Because epistemic probability for Bernoulli is more nearly parallel to expectation than to probability in the modern frequentist sense, there can be alternative probabilities that add up to more or less than one (see Shafer 1978). As Bernoulli explains:

> Here it should be noted that, if the arguments adduced on each side are strong enough, it may happen that the absolute probability of each side significantly exceeds half of certainty, that is, that both of the contraries are rendered probable, though relatively speaking one is less probable than the other. Thus it can happen that one thing has 2/3 of certainty, while its contrary has ¾; in this way both contraries will be probable, yet the first less probable than its contrary in the ratio 2/3 to ¾ or 8 to 9.
>
> (Bernoulli 2006: p. 325; Bernoulli 1713/1968: p. 221).

It is worth noting, then, that not only are Parts I–III of *Ars Conjectandi* not based on probability as relative frequency, but that even Part IV, which does concern probability so-named, is largely concerned with epistemic and not frequentist probability.

3.6 Bernoulli's Philosophical Outlook

In Bernoulli's conception, the art of conjecture would lead to probable and not certain guides to decision-making. Before describing the mathematics of the art of conjecturing, Bernoulli stated nine 'general rules or axioms, which simple reason commonly suggests to a person of sound mind, and which the more prudent constantly observe in civil life' (Bernoulli 2006: p. 318; Bernoulli 1713/1968: p. 214). The last two of these are:

> 8. In our judgments we should be careful not to attribute more weight to things than they have. Nor should we consider something that is more probable than its alternatives to be absolutely certain, or force it on others.
>
> 9. Because, however, it is rarely possible to obtain certainty that is complete in every respect, necessity and use ordain that what is only morally certain be taken as absolutely certain.
>
> (Bernoulli 2006: pp. 320–21; Bernoulli 1713: pp. 216–17)

It is repeated in several sources that by his theorem Bernoulli sought to achieve 'moral certainty' (e.g. Stigler 1986: p. 65). It is true that Bernoulli defined 'morally certain' by saying:

> Something is *morally certain* if its probability comes so close to complete certainty that the difference cannot be perceived. ....if we take something that possesses 999/1000 of certainty to be morally certain, then something that has only 1/1000 of certainty will be morally impossible.
>
> (Bernoulli 2006: p. 316; Bernoulli 1713/1968: pp. 211–12)

It is also true that in choosing numbers for an example of the results of his theorem Bernoulli chose a probability of 1000 to 1 that the observed ratio falls between the ratios 29:50 and 31:50, but he did not here call a probability of 1000 to 1 morally certain. Rather, moral certainty was to be chosen institutionally. In commenting on his ninth rule or axiom, Bernoulli wrote:

> It would be useful, accordingly, if definite limits for moral certainty were established by the authority of the magistracy. For instance, it might be determined whether 99/100 of certainty suffices or whether 999/1000 is required. Then a judge would not be able to favor one side, but would have a reference point to keep constantly in mind in pronouncing a judgment.
>
> (Bernoulli 2006: p. 321; Bernoulli 1713/1968: p. 217)

Had Bernoulli tried to find a ratio of outcomes a posteriori in practice, he would not have demanded a probability as high as 1000 to 1 unless the stakes were extremely high.

Earlier, when Bernoulli distinguished between necessity and contingency in Part IV of *Ars Conjectandi* – where 'probable' had traditionally meant that the chances of occurrence were (well) over half – he distinguished three senses of 'necessary':

> Something is *necessary* if it cannot not exist, now, in the future, or in the past. This necessity may be physical, hypothetical, or contractual. It is *physically necessary* that fire burn, that a triangle have three angles equal to two right angles, and that a full moon occurring when the moon is at a node be eclipsed. It is *hypothetically necessary* that something, while it exists or has existed, or while it is assumed to exist or have existed, cannot not exist or not have existed. It is necessary in this sense that Peter, whom I know and posit to be writing, is writing. Finally, there is the *contractual* or *institutional necessity* by which a gambler who has thrown a six is said to win necessarily if the players have agreed beforehand that a throw of six wins.
>
> (Bernoulli 2006: p. 316; Bernoulli 1713/1968: p. 212)

When Huygens and Bernoulli calculate expectation, it is contractual or institutional necessity (i.e. the players' agreement to abide by the rules of the game) that determines the necessity. Then when Bernoulli turns to the probability of conjectures, it is still more often contractual than physical necessity that determines the probability or certainty.

When *Ars Conjectandi* was published in 1713, the volume also included Bernoulli's *Lettre à un Amy sur les Parties du Jeu de Paume*. In his letter, Bernoulli calculated the handicaps that might be given by one player to another in games of court tennis to make their chances of winning equal. He asserted that the relative strengths of players could be learned by watching the strokes each player won against the other over a large number of strokes (Bernoulli 2006: pp. 263–4). Bernoulli proposes in this connection that observing

the outcomes of 300 or 400 strokes would be enough to estimate a posteriori the relative strengths of two players. This is far less than the 25,550 observations that he calculated would be needed to determine with a probability of 1000 to 1 that a ratio of outcomes fell between 29/50 and 31/50 (Bernoulli 2006: pp. 337–9; Sylla 2013). This is clearly because the issue has to do only with one friend giving another a handicap or advantage at the beginning of a game so that they would have an approximately equal chance of winning the game. Not only would it be unrealistic to wish to observe more than 25,000 rallies between the two players, but exactness was not appropriate in the situation.

3.7 Frequentist Probability in the Works of Abraham De Moivre

Whereas Bernoulli combined epistemic probability in the art of conjecturing with the mathematics of expectations in games of chance, Abraham De Moivre more clearly shifted the core meaning of probability toward relative frequency. Already in the title of his 1712 work, *De mensura sortis, seu, de Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus*, De Moivre wrote about probability in a way that can be understood in a frequentist sense (See Hald 1984: pp. 237–8). In much of *De mensura sortis*, however, De Moivre used 'probable' to mean 'with a likelihood of more than 50 percent'.

Then, in the first edition of his *Doctrine of Chances* (1718), which had the subtitle: *or, a Method of Calculating the Probabilities of Events in Play*, De Moivre noted up front that the sum of probabilities in a given situation must equal 1:

> The Fractions which represent the Probabilities of happening and failing, being added together, their Sum will always be equal to Unity.
>
> (De Moivre 1718: p. 2)

By the third edition of *The Doctrine of Chances* (1756), De Moivre elaborates:

> ….it being a certainty that an Event will either happen or fail, it follows that Certainty, which may be conceived under the notion of an infinitely great degree of Probability, is fitly represented by Unity.
>
> These things will easily be apprehended, if it be considered, that the word Probability includes a double Idea; first, of the number of Chances whereby an Event may happen; secondly, of the number of Chances whereby it may either happen or fail …. It is the comparative magnitude of the number of Chances to happen, in respect to the whole number of Chances either to happen or to fail, which is the true measure of Probability.
>
> (De Moivre 1718/1756: p. 2)

Immediately after this, De Moivre turns to expectation and its measure:

> In all cases, the Expectation of obtaining any Sum is estimated by multiplying the value of the Sum expected by the Fraction which represents the Probability of obtaining it.
>
> (De Moivre 1718/1756: p. 3; except for 'the Fraction which represents' this was already in the 1718 edition)

Thus De Moivre clearly differentiates probability as relative frequency from expectation, thereby making it possible to define expectation as the value to be obtained times the probability of obtaining it.

3.8 Summary and Conclusion

To sum up: in Bernoulli's *Ars Conjectandi* the concept of 'lot' or 'expectation' in games, as found in Huygens' *De ratiociniis* and earlier works, was linked, through a transfer of mathematics from one conceptual realm to another, to the epistemic probability of opinions. Then in the work of De Moivre the concept of 'lot' or 'expectation' was unpacked so that it was now understood more clearly as the product of the value one might expect times the relative frequency or likelihood (now labeled 'probability') of receiving it. In this chapter, I have tried to distinguish Jacob Bernoulli's concepts from those of Huygens on the one hand and those of De Moivre on the other, so that it might be possible to detect how slowly the concepts related to modern mathematical probability came together, the conceptual development trailing the mathematical development in some cases.

The emergence of mathematical probability in the seventeenth and eighteenth centuries was the work of mathematicians. The histories of this period in Anders Hald 1990 and 1998 are excellent, and I have written my essay assuming that the reader can look for the rest of the mathematical story in Hald's books. A short contextual history of the origin of mathematical probability may be found in Callinger 1999: pp. 548–54, 649–54.

Bernoulli argued that mathematics can play an important role in promoting clear reasoning. He refers to Pascal's first letter to Fermat to make this point. Pascal had mentioned 'a certain anonymous man, otherwise of cultivated judgment, but devoid of mathematics' – identified as Antoine Gombault de Meré – who, according to Pascal, was perplexed because there seemed to be inconsistencies in calculations of the numbers of throws needed to have more than a fifty percent chance of winning a bet in various dice throws (Bernoulli 2006: p. 157; Bernoulli 1713/1968: p. 29). From Bernoulli's perspective, Pascal's letters are evidence of how mistaken laymen may be in their reasoning if they do not make use of mathematics.

Here, from the perspective of Jacob Bernoulli's *Ars Conjectandi*, I have tried to look at the concepts by which mathematicians such as Huygens, Bernoulli, and De Moivre linked mathematical techniques to the subject matters of expectation in games of chance and then to probability in its epistemic or frequentist senses. Had I more space, I would have said more about G. W. Leibniz, who later claimed that Jacob Bernoulli had pursued the mathematics of probability with his encouragement (see Sylla 1998, 2006). Leibniz was capable of making elementary mistakes in mathematical probability such as saying that with two dice it is equally difficult to throw 11 and 12 (Hacking 1975: pp. 52–3), but he wrote a huge amount of work related to mathematical probability, most of it unpublished in his lifetime. For evidence of this one need only look at the collections of fragments in Parmentier (1995) and Leibniz (ed. Knobloch 2000), together with the article by Schneider (2000) in the latter volume.

Jacob Bernoulli was risk averse, saying that the art of conjecturing was:

> the art of measuring the probabilities of things as exactly as possible, to the end that, in our judgments and actions, we may always choose or follow that which has been found to be better, more satisfactory, safer, or more carefully considered.
>
> (Bernoulli 2006: pp. 317–18)

He was not attempting to help his readers maximize their winnings or profit without taking account of risk. Bernoulli's rules or axioms for the art of conjecturing were sober or conservative. So his fifth rule was:

> In matters that are uncertain and open to doubt, we should suspend our actions until we learn more. But if the occasion for action brooks no delay, then between two actions we should always choose the one that seems more appropriate, safer, more carefully considered, or more probable, even if neither action is such in a positive sense.
>
> (Bernoulli 2006: pp. 318–21; Bernoulli 1713/1968: pp. 214–17)

After the mathematical work of Bernoulli and De Moivre, many philosophical issues remained open. In the later eighteenth century, Moses Mendelssohn attempted to apply the work of the mathematical probabilists to the problems of induction, determinism, and God's knowledge (see Sylla 2011). Bernoulli's proof of his fundamental theorem essentially treats all the possible outcomes of trials as simultaneous, in effect taking a 'God's eye' view, God being eternal and outside of time (See Sylla 2006b). In the gamble at the center of the St. Petersburg paradox, about which Jacob's nephew Daniel Bernoulli would write, the value of the expectation holds only if the player can play *ad infinitum*. To resolve the apparent paradox, Daniel proposed the concept of (marginal) utility.

Similarly, the standard solution to the division problem assumes that players are indifferent regarding whether to continue with the game, on the one hand, or leave after having been paid what their expectations are at that stage of the game, on the other (cf. Pascal 1665/1970: p. 1308). This, however, does not take account of differences in the risk tolerance of different individuals. The Port Royal logic had earlier made the argument that in making decisions about actions, one should consider not only the benefit or harm that might result from a given action, but also the relative frequency with which various results occur. For example, one's fear of being struck by lightning, however awful the outcome might be, should be tempered by the rarity of such events (Sylla 2006: p. 25 and 25 n.). The establishment of expectation in the sense of the mathematical product of benefit or cost times relative likelihood (or the combination of several such expectations into a single sum) as a significant magnitude did not resolve all the disagreements that might arise in the future over cost-benefit analyses.

References

Bellhouse, D. R. (2005) Decoding Cardano's *Liber de Ludo Aleae*. *Historia Mathematica*. 32. pp. 180–202.

Bernoulli, Jacob (1744/1967) *Opera*. (Sumptibus Haeredum Cramer et Fratrum Philibert). Repr. Brussels: Culture et Civilisation.

Bernoulli, Jacob (1713/1968) *Ars Conjectandi*. Opus posthumum. Accedit *Tractatis de seriebus infinitis,* et epistola Gallicè scripta *De ludo pilae reticularis*. Repr. Brussels: Culture et Civilisation.

Bernoulli, Jacob (1975) *Die Werke von Jakob Bernoulli*. (ed. B. L. van der Waerden.) Basel: Birkhäuser.

Bernoulli, Jacob (1993) *Der Briefwechsel von Jacob Bernoulli*. Basel: Birkhäuser.

Bernoulli, Jacob (2006) *The Art of Conjecturing* and *Letter of a Friend on Sets in Court Tennis* Translated from the Latin by Edith Dudley Sylla. Baltimore, MD: Johns Hopkins University Press.

Bernoulli, Nicolaus (1709/1975) *De Usu Artis Conjectandi in Jure*. In Bernoulli, Jacob *Werke*. pp. 287–326. Basel: Birkhauser.

Bernoulli, Nicolaus (1711) Specimina Artis Conjectandi, ad quaestiones Juris Applicata. *Acta Eruditorum*. 4. 4. pp. 159–70.

Callinger, R. (1999) *A Contextual History of Mathematics to Euler*. Upper Saddle River, NJ: Prentice Hall.

Cardano, G. (1539/1662) *Practica arithmetice*. In *Opera Omnia*. Vol. 4. Lyon: Ioannes Antonius Huguetan and Marcus Antonius Ravaud. (First published in Milan.)

Daston, L. (1988) *Classical Probability in the Enlightenment*. Princeton, NJ: Princeton University Press.

De Moivre, A. (1711) De Mensura Sortis, seu, De Probabilitate Eventuum in Ludis a Casu Fortuito Pendentibus. *Philosophical Transactions*. 329. (January, February, March). pp. 215–64.

De Moivre, A. (1718) *The Doctrine of Chances, or, A Method of Calculating the Probability of Events in Play*. [2^{nd} ed. 1738; 3^{rd} ed. 1756.] London: W. Pearson.

Edwards, A. W. F. (2002) *Pascal's Arithmetic Triangle*. 2^{nd} ed. Baltimore, MD: The Johns Hopkins University Press.

Edwards, A. W. F. (2003) Pascal's Work on Probability. In Hammond, N. (ed.) *The Cambridge Companion to Pascal*. Cambridge: Cambridge University Press.

Franklin, J. (2001) *The Science of Conjecture. Evidence and Probability before Pascal*. Baltimore, MD: The Johns Hopkins University Press.

Hacking, I. (1975) *The Emergence of Probability*. Cambridge: Cambridge University Press.

Hald, Anders (1984) A. De Moivre: 'De Mensura Sortis' or 'On the Measurement of Chance'. *International Statistical Review*. 52. 3. pp. 229–62.

Hald, Anders (1990) *A History of Probability and Statistics and Their Applications before 1750*. Hoboken, NJ: John Wiley.

Hald, Anders (1998) *A History of Mathematical Statistics from 1750 to 1930*. Hoboken, NJ: John Wiley.

Huygens, Christiaan (1657) *De ratiociniis in ludo aleae*. Included in Jacob Bernoulli 1713 and translated in Jacob Bernoulli 2006. Brussels: Culture et Civilisation and Baltimore, MD: The Johns Hopkins University Press.

Leibniz, G. W. (2000) *Hauptschriften zur Versicherungs- und Finanzmathematik*. Berlin: Akademie Verlag.

Meusnier, N. and Piron, S. (2007) Medieval Probabilities: A Reappraisal. [See also other articles.] Journal Electronique d'Histoire des Probabilités et de la Statistique / Electronic Journal for History of Probability and Statistics. 3 (1). pp. 1–14 [Online] Available from: <www.jehps.net> [Accessed 7 Oct 2015.]

Montmort, Pierre Rémond de (1708) *Essay d'analyse sur les jeux de hazard*. Paris: Chez Jacque Quillau, Imprimeur-Juré-Libraire de l'Université.

Paccioli, Fra Luca (1494) *Summa de Arithmetica geometria proportioni & proportionalita*. Venice: Paganino de Paganini.

Parmentier, M. (1995) *G. W. Leibniz, l'estime des apparences. 21 manuscrits de Leibniz sur les probailités, la théorie des jeux, l'espérence de vie*. Paris: Vrin.

Pascal, Blaise (1665/1970) *Traité du Triangle Arithmétique*. Included in Pascal, *Oeuvres Complètes de Blaise Pascal*, Vol. II. pp. 1166–332.

Pascal, Blaise (1970) *Oeuvres Complètes de Blaise Pascal* Vol. II. 1623–54. Bruges: Desclée De Brouwer.

Schneider, Ivo (2000) Geschichtlicher Hintergrund und wissenschaftlichen Umfeld der Schriften. In Leibniz *Hauptschriften zur Versicherungs- und Finanzmathematik*. pp. 591–623. Berlin: Akademie Verlag.

Shafer, Glenn (1978) Non-Additive Probabilities in the Work of Bernoulli and Lambert. *Archive for History of Exact Sciences*. 19. pp. 309–70.

Stigler, Stephen (1986) *The History of Statistics. The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press of Harvard University Press.

Sylla, E. D. (1990) Political, Moral, and Economic Decisions and the Origins of the Mathematical Theory of Probability: The Case of Jacob Bernoulli's. *The Art of Conjecturing*. In von Furstenburg, G. (ed.) *Acting Under Uncertainty: Multidisciplinary Conceptions*. pp. 19–44. Dordrecht: Kluwer).

Sylla, E. D. (1997) Jacob Bernoulli on Analysis, Synthesis, and the Law of Large Numbers. In Otte, M. and Panza, M. (eds.) *Analysis and Synthesis in Mathematics: History and Philosophy*. pp. 79–101. Dordrecht: Kluwer.

Sylla, E. D. (1998) The Emergence of Mathematical Probability from the Perspective of the Leibniz-Jacob Bernoulli Correspondence. *Perspectives in the Sciences.* 6. pp. 41–76.

Sylla, E. D. (2003) Business Ethics, Commercial Mathematics, and the Origins of Mathematical Probability. In Schabas, M. and De Marchi, N. (eds.) *Oeconomies in the Age of Newton*. Annual Supplement to *History of Political Economy.* 35. pp. 309–27. Durham, NC: Duke University Press.

Sylla, E. D. (2006a) Introduction and notes. In Jacob Bernoulli. *The Art of Conjecturing and Letter of a Friend on Sets in Court Tennis*. Translated from the Latin by Edith Dudley Sylla. Baltimore, MD: Johns Hopkins University Press.

Sylla, E. D. (2006b) Commercial Arithmetic, Theology and the Intellectual Foundations of Jacob Bernoulli's *Art of Conjecturing*. In Poitras, G. (ed.) *Pioneers of Financial Economics* Vol. 1 *Contributions Prior to Irving Fisher*. Cheltenham: Edward Elgar Publishing. (Revised and expanded version of Sylla 2003).

Sylla, E. D. (2011) Mendelssohn, Wolff, and Bernoulli on Probability. In Munk, R. (ed.) *Moses Mendelssohn's Metaphysics and Aesthetics*. pp. 41–63. Berlin: Springer.

Sylla, E. D. (2013) Jacob Bernoulli and the Mathematics of Tennis. *Nuncius*. 28. pp. 142–63.

Sylla, E. D. (2014) Tercentenary of *Ars Conjectandi* (1713): Jacob Bernoulli and the Founding of Mathematical Probability. *International Statistical Review*. 82. pp. 27–45.

van der Waerden, B. L. (1975) Introduction and notes. In *Die Werke von Jakob Bernoulli*. Bd. 3. Basel: Birkhäuser.

Yushkevich, A. P. (1987) Nicholas Bernoulli and the Publication of James Bernoulli's *Ars Conjectandi*. *Theory of Probability and its Applications*. 31. 2. pp. 286–303.

CHAPTER 4

PROBABILITY AND ITS APPLICATION IN BRITAIN DURING THE 17TH AND 18TH CENTURIES

DAVID R. BELLHOUSE

4.1 Introduction

Abraham De Moivre reviewed a book by Ludwig Martin Kahle for the Royal Society at the beginning of 1736 (see Bellhouse 2011: pp. 117–18). At the time, De Moivre was working on the second edition of his *Doctrine of Chances*. Written in Latin, Kahle's (1735) *Elementa Logicae Probabilium* covered ideas in probability, both ancient and modern. In his review De Moivre commented in a way that touches on the development of probability in Britain:

> This doctrine was first introduced by Huygens in the year 1657, in a little Treatise intitled: Ratiocinia de Ludo Aleae: and has since been followed by James Bernoulli, Monmort, Nicholas Bernoulli, myself and perhaps others. For altho' the ancients had the same Idea of Probability as we have; and they mentioned several degrees of it, as appears in the words Probabile and Probability often used by Cicero & other writers, yet the Author [Kahle] observes, that the distinct measure of it was never assigned till the times above mentioned: which occasions in his Book an entertaining dissertation full of Learning and polite literature.

Among the personalities that De Moivre mentioned as major contributors to the development of probability in his day, the only one from Britain was De Moivre himself. That was not arrogance or boasting, but a statement of fact. He was a pivotal figure to the extent that probability in Britain may be divided into two major epochs: before De Moivre and after De Moivre, with the dividing point given by the publication of De Moivre's *De Mensura Sortis* in 1711 (De Moivre 1711), his first venture into probability theory.

The events before 1711 might be characterized in the main by a theory espoused by Huygens looking for applications. The early applications made, described in §2, were to

politics (Richard Cumberland in response to Thomas Hobbes), gambling (Isaac Newton's response to Samuel Pepys on a dicing question), medicine (Archibald Pitcairne on the circulation of blood) and theology (John Craig on faith in the Christian story). Minor developments in the mathematical theory also took place (Isaac Newton and Thomas Strodes). After 1711 there was a major switch in emphasis to solving mathematically difficult problems in probability and to finding efficient ways to calculate "the distinct measure" of probability. There was not a distinct change point in 1711; rather there was a transitional period just prior to that time, which is described in §3. The precursors of change, John Arbuthnot and Francis Robartes, were virtuoso amateur mathematicians, not professional ones. Arbuthnot was a physician and satirical writer, while Robartes was a politician with roots in the aristocracy. Abraham De Moivre was one of the best mathematicians of the eighteenth century in Britain. His work in the theory of probability is described in §4 and his major application of probability, the valuation of life annuities, is described in §5, along with the work of his major rival in this area, Thomas Simpson. Although De Moivre broke new paths in the theory of probability, he also in a sense confined it. By 1718 he had decided that all problems in probability could be solved using the binomial theorem and infinite series (De Moivre 1718: p. viii). This became the theme for his theoretical work and for many who followed him. Towards the end of this era, as described in §6, Thomas Bayes broke new ground with his theorem related to inferences about future events given past experience. It was brought to light by his friend, Richard Price, probably because of the philosophical and theological implications of Bayes's mathematical results. Bayes's work was, however, ignored in Britain even into the early nineteenth century.

4.2 The Impact of Christiaan Huygens: A Theory in Search of Applications

De Moivre had read Huygens's (1657) *De Ratiociniis in Ludo Aleae* in the early 1680s when he was a student in France (Maty 1755; Bellhouse and Genest 2007: p. 111). As the Latin title suggests, Huygens's probability problems were all set in terms of games of chance (literally *ludi aleae*). Despite the title, the book was not aimed at gamblers. It was part of a book of mathematical exercises. What Huygens had done was to solve in a systematic way two long-standing yet current mathematical problems: the problem of the division of stakes and problems related to the sum of the faces that show in the throw of dice. The division of stakes problem appears in several Renaissance Italian commercial arithmetic books. Two players play a series of games and the first to win a predetermined number of them wins the pot. Partway into the series, after each player has won some of the games (perhaps not the same number), it is decided to terminate the series and split the pot between the players. How should the pot be split? Hald (1990: pp. 34–6) describes some of the unsuccessful attempts by Italian Renaissance mathematicians at solving this problem. The earliest known statement of the problem of the sum of the faces that show on dice is contained in a late medieval morality poem, *De Vetula.* In the poem there is a long and cumbersome enumeration of the chances of every possible sum in the throw of three dice. Bellhouse (2000) describes in detail the material in the poem related to the dicing problem.

Huygens's general approach to solving probability problems was to attach a value or payout for each of the outcomes in a game of chance and then to find the expected payout rather than the probability. When calculating an expectation, Huygens (1657) continually refers to the number of chances to win or to lose, implying that these chances can be enumerated. His approach was influential and was followed extensively by Pierre Rémond de Montmort (Montmort 1708 and 1713) in his *Essay d'analyse sur les jeux de hazard* and to a lesser extent by Jakob Bernoulli (Bernoulli 2006) in his *Ars Conjectandi*, published posthumously in 1713.

Isaac Newton read *De Ratiociniis in Ludo Aleae* as a student at Cambridge and later made some probability calculations at the urging of Samuel Pepys.

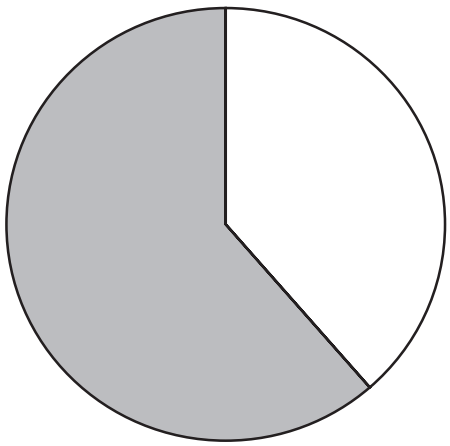


FIGURE 4.1

Newton's infrequent dabbling in probability over a period of nearly thirty years shows the spectrum of approaches to the calculation of what De Moivre had called "the distinct measure" of probability. In about 1665, while still a student, Newton reproduced some of Huygens's results and expanded on them slightly (Newton 1967: pp. 58–62). While Huygens considers chances to win or lose as whole numbers, Newton provides an example to show how the concept can be extended. As an example that is similar to what Newton did, consider the circle in Figure 1 that is divided into a grey and a white part by two radii shown in the circle. If a ball is dropped onto the centre of the circle then the expectation for any bet on where the ball will fall is the weighted average of the possible amounts won, with the weights given by the relative sizes of the grey and white areas. Newton uses the specific example of 2 and $\sqrt{5}$ for the values of the sizes. Later in 1693, Newton was presented with a probability problem by Samuel Pepys (see Stigler 2006 and Pepys 1926). The problem is to determine which of three events has the greatest probability of occurrence: (A) throwing at least one six in the throw of six dice; (B) throwing at least two sixes in the throw of twelve dice; or (C) throwing at least three sixes in the throw of eighteen dice. The modern approach to solving the problem is to recognize that the number of sixes that show in each situation follows a binomial distribution with probability of success given by 1/6 and then to calculate the appropriate binomial probabilities. This yields, to six decimal places, the probabilities:

(A) 0.665102; (B) 0.6186674; and (C) 0.5973457. Newton first used a fallacious logical argument without any numerical calculation to conclude correctly that A has the best chance followed by B and then C. When comparing A and B, he reasoned that all the possibilities of getting a six are present for A when throwing six dice. For B, however, they are not; if only a single six appears in the throw of twelve dice, he would not win. Similarly when C enters the picture, if one or two sixes appear in the throw of eighteen dice, he would not win. Consequently, A's chances are better than B's, which are better than C's. Stigler (2006) comments, "Newton's proof refers only to the sample space and makes no use of the probabilities of different outcomes other than that the dice are thrown independently, and so it must fail." Using weighted dice, Stigler provides a counterexample to Newton's proof. After prodding from Pepys for a numerical answer, Newton did an exact enumeration in each case of the number of favourable outcomes and the total number of outcomes. His method of enumeration followed neither Huygens's approach nor De Moivre's preferred approach through the use of the binomial theorem.

After Newton, a diverse group in Britain was influenced by Huygens's work. The group includes two physicians from Scotland, Archibald Pitcairne and John Arbuthnot, as well as an Anglican clergyman, Richard Cumberland, who became Bishop of Peterborough. Another Englishman, Francis Robartes, younger son of the Earl of Radnor, reacted negatively to Huygens's approach to the calculation of probabilities. There appears to be no relation between the efforts made by all of these individuals, with the one connecting thread being Huygens.

Cumberland used probability results from Huygens as one small building block in an argument he made against Thomas Hobbes's 1651 *Leviathan* (Hobbes 1994). In *Leviathan*, Hobbes argues, using scientific and mathematical premises, that people should submit themselves to the authority of an absolute sovereign. Hobbes contends that force or fraud is the natural road to personal security and, consequently, that there are no unjust actions (Hobbes 1994: p. 75). He argues for a strong government in order to avoid this natural state of mankind. Hobbes also argues that "self-interest lies at the heart of political and ethical theory" (Parkin 1999: p. 6) and he separates science from ethics. Many of his contemporaries labeled him an atheist. Cumberland's response to Hobbes was made in 1672 in his *De Legibus Naturae* (Cumberland 2005). A major motivation for Cumberland to write was that acceptance of Hobbes's claims in *Leviathan* would substantially undermine the spiritual and temporal power of the Church of England, to which he was deeply committed. As just one small building block in his argument against Hobbes, Cumberland relies on Huygens's *De Ratiociniis* to make a point (see Stigler 1999: pp. 239–51). From Huygens, he notes that one can find the value or expectation of outcomes that are uncertain. Cumberland goes on to say that events that are more probable will have greater value in expectation. By using some dicing examples from Huygens, Cumberland (2005: pp. 492–3, 625, 643–4) argues that unjust men are more likely to be exposed to danger, and therefore natural punishments, than just men. This refuted, Cumberland thought, Hobbes's claims of the natural benefits of force or fraud.

Pitcairne was interested in the mechanics of the circulation of blood through the body (Friesen 2003: p. 169). He believed that the cause of illness was impaired circulation and he wanted to subject the hydraulics of blood circulation to mathematical study. Previously, Harvey (1628) had put forward the idea of the circulation of blood by means of the heart as a pump. By Pitcairne's time, it was thought that blood was composed of a mixture of

particle-based fluids that were each secreted to different parts of the body through sieve-like openings, or pores, in various glands. The particles could be of similar shape (homogeneous) or of different shapes (heterogeneous) and the holes in the sieve would control which part of the body would receive a particular particle of blood. He then examines the angles at which either kind of particle can pass through the sieve. Using probability arguments from Huygens, Pitcairne (1693) studies this hypothetical model. He bases his argument on Proposition III in Huygens (1657). In an English translation by Edith Dudley Sylla (Bernoulli 2006: p. 135) the proposition reads:

> If the number of cases in which a falls to me is p and the number of cases in which b falls to me is q, and if all the cases can happen equally easily, then my expectation will be worth $(pa + qb)/(p + q)$.

Rather than an amount won, Pitcairne lets b be the conditions for admission through the sieve and a the conditions for exclusion. Pitcairne sets p to be the number of angles for which a particle is excluded and q to be the number of angles in which admission can occur. For homogeneous particles, he argues that p is finite and q is infinite so that the expectation $(pa + qb)(p + q)$ reduces to b. Consequently, admission through a pore always occurs. For heterogeneous particles, the opposite is true – p is infinite and q is finite – so that the expectation is a and exclusion always occurs. He could not find a good alternative to explain how secretion of blood would occur, but could only argue that the pores were necessarily circular in shape (see Stigler 1999: pp. 203–38). Pitcairne's arguments might be questionable but his reasoning does show the influence of Huygens's work in probability.

There were a few from this period who were not influenced by Huygens. In 1678, Thomas Strode published a treatise on permutations and combinations (Strode 1678). He uses one of his results in this area to find the number of outcomes of the faces that show on up to six dice. For two, three, and four dice, Strode enumerates all possible throws and uses his counting methods to find the number of chances for each of the possible sums of the faces that show. Then he shows how his methods can be generalized to dice that have other than six faces. The only connection to Huygens is that Huygens had solved the problem of the sum of the faces that show in the throw of three dice while Strode extends the result to a few more than three dice (see Stigler 1999: pp. 239–51). Much farther removed from Huygens is John Craig, a Church of England clergyman and mathematician who made probability calculations that are not probabilities strictly speaking; because the probability measures he used are not each between 0 and 1 and do not sum or integrate to 1. Influenced by Newton's *Principia Mathematica*, the calculations appear in Craig's 1699 *Theologiae Christianae Principia Mathematica* (Nash 1991). Craig's major application of his theory is an assessment of belief in the story of Jesus of Nazareth and when faith in this story would disappear. With the exception of Stigler (1999: pp. 252–76), who has tried to rescue this work from the ashcan of abuse by giving it a sympathetic analysis, most writers from the nineteenth century on have been very critical of Craig's work. It is difficult to rescue Craig's approach to probability in the context of the developments in probability up to his time or even today. It is much easier to describe his result and to understand it in the context of trying to adapt ideas from Newton's monumental and highly influential work. Craig's analysis was driven by the science and politics, intertwined with the prevailing religious beliefs, of his day.

4.3 A Time of Transition: John Arbuthnot and Francis Robartes

Archibald Pitcairne was a friend of the mathematician David Gregory from the time they met at the University of Edinburgh. Pitcairne and Gregory gathered round themselves a circle of former students, all Scots (Guerrini 1986). Included in this group was John Craig. Another member was the physician, satirist, and mathematician John Arbuthnot.

Born in Scotland, Arbuthnot went south to London in search of fame and fortune. Shortly after his arrival in London, Arbuthnot published a translation of Huygens's *De Ratiociniis*, which he entitled *Of the Laws of Chance* (Arbuthnot 1692). This was not an academic exercise that would put Huygens's Latin text into the hands of English mathematics students. Admittedly, at first glance it does appear framed that way. When examined carefully, however, it may be viewed as the first gambling manual in English to contain probability calculations that the author hoped would be of benefit to those playing games of chance. *Of the Laws of Chance* falls in line with a substantial literature on gambling in England. The earliest literature of this genre, beginning in the sixteenth century, is concerned mainly with descriptions of how to detect cheating at cards and dice done, for example, through the use of loaded dice and marked cards. Beginning in the mid-seventeenth century, rulebooks appear that are descriptions of how to play various games. The theme is the same throughout this entire literature: educate the unwary about both the rules and methods of cheating so that cheaters cannot take advantage of the ordinary person playing the game (see Bellhouse 1993). With Arbuthnot's book, a new twist is added; further advice on games is given through probability calculations.

Approximately the first half of Arbuthnot's text is a faithful translation of Huygens's *De Ratiociniis* written in a very fine literary style. His style and faithfulness can be seen by comparing the very first sentence of his translation to a modern and very literal translation done by Edith Dudley Sylla (Bernoulli 2006).

Arbuthnot	Sylla
"Although the Events of Games, which Fortune solely governs, are uncertain, yet it may be certainly determin'd, how much one is more ready to lose than gain."	"Although the outcomes of games that are governed purely by lot are uncertain, the extent to which a person is closer to winning than to losing always has a determination."

In the second half of Arbuthnot's book, several games of chance are considered and the probabilities of various types of play in the games are determined. Very little detail about some of the games is given; Arbuthnot must have assumed that the reader was familiar with these games. In the new material Arbuthnot is faithful to Huygens's approach by using expected values or proportions of the stake throughout. Arbuthnot mentions five games by name: Backgammon, Hazard, Raffle, Royal Oak Lottery, and Whist. With the exception of Backgammon, De Moivre (1718) examines all these games in much greater detail in his *Doctrine of Chances*.

Arbuthnot is best known for his work on the sex ratio, published in 1710 (Arbuthnott 1710). Using what is probably the first published test of significance, Arbuthnot argues for divine providence, or God's actions in the world. The data for the test are taken from the London Bills of Mortality, collected and published by the Company of Parish Clerks of London. For 82 years in a row, from 1629 to 1710, more males than females were born in London. Arbuthnot had noticed the same result in 1692 in his *Of the Laws of Chance* but with fewer data points. In 1710, based on his assumption that male and female births occur by chance, by which he means they are equiprobable, Arbuthnot calculates the probability of obtaining what had been observed in London for 82 years. The resulting probability is $1/2^{82}$, a very small number. To Arbuthnot, this was evidence of divine providence. Attempts at this kind of test are found in a manuscript Arbuthnot wrote in 1694 (see Bellhouse 1989), where he tries not very successfully to evaluate the validity of the chronologies of the first seven kings of Rome and the 111 kings of Scotland.

In the 1694 manuscript and the 1710 paper, Arbuthnot introduces the binomial theorem to solve some probability problems. His use of the binomial theorem may have come about from his membership in the Pitcairne-Gregory circle. It was Pitcairne who gave the first printed version of the general expression of the binomial theorem in 1688 and attributed the result to Gregory (see Stigler 1999: pp. 203–38).

Arbuthnot uses the binomial expansion in 1694 to solve in general a question posed by Huygens in 1657: 'How many pairs of dice does one need to throw in order that the probability of getting at least two sixes is at least $\frac{1}{2}$?' Arbuthnot's generalization is to consider a die with f faces thrown n times and then to find the value of n such that the probability of getting at least one specified side is exactly some given amount. In modern probability jargon, the problem is to find how many trials n are required in a series of independent Bernoulli trials, each with probability of success $1/f$, so that the probability of obtaining at least one success is p.

In the 1710 paper Arbuthnot uses another argument about the sex ratio to try to explain why male and female births do not occur with equal frequency; the same argument appears with more detail in the 1694 manuscript. The argument is based on the fact that for a large number of births the probability of obtaining exactly the same number of male and female births is very unlikely. He begins by considering flipping a coin a set number of times and then finds the probability of obtaining exactly an equal number of heads and tails. In the 1694 manuscript he considers 1,000,000 tosses, for which the probability is

$$\frac{1000000 \times 999999 \times 999998 \times \cdots \times 500002 \times 500001}{2^{1000000} \times 500000 \times 499999 \times 499998 \times \cdots \times 3 \times 2 \times 1}$$

to obtain an equal number of heads and tails. Arbuthnot has trouble evaluating or even approximating this value. Using an argument from logarithms, he thinks that his probability is in the order of $10^{-251030}$. Consequently he argues, both in 1694 and in 1710, that the equilibrium that is kept between the sexes cannot be due to chance.

In what he perhaps considered as the crowning achievement of his career in probability, De Moivre later obtained the correct general expression for Arbuthnot's approximation to the middle term of the binomial. As applied to Arburthnot's problem, the probability is actually close to 8×10^{-4}.

Francis Robartes, another gentleman virtuoso mathematician, did not accept at face value what Huygens had written in *De Ratiociniis*. Late in the seventeenth century Robartes

(Roberts 1693) thought he detected a probabilistic paradox and published it in the *Philosophical Transactions*. Robartes's work has generally been ignored probably because of Isaac Todhunter's assessment of it in his *History of the Theory of Probability*. Todhunter (1865: pp. 54–5) sums up the paper with:

> "The paradox is made by Roberts [Robartes] himself, by his own arbitrary definition of *odds*."

On the contrary, this was not an arbitrary definition but one taken from reading Huygens. Take, for example, the outcome of throwing at least one six in the throw of two dice. Using Huygens's approach, for a stake of value a, the expectation to the player betting on this event is $11a/36$ so that the expectation to his opponent is $25a/36$. Arbuthnot's 1692 translation of Huygens's solution to the problem states that the bet of the first player

> "is worth $11a/36$, so there remains to his Fellow-Gamester $25a/36$; so the Value of my Expectation to his, is as 11 to 25, *i.e.* less than 1 to 2."

The term "11 to 25" also refers to the odds in favour of the player's winning. Huygens does this in a few other places – taking the ratio of the expectations to obtain an expression that provides an equivalent and correct expression for the odds of winning. Robartes makes the leap that generally the odds of winning are the ratio of the total gains that could be obtained by each player. This is in the spirit of Huygens and it is central to Robartes's paper. What Robartes does is argue against this general leap by constructing a simple counterexample to show that the ratio of the expectations is not necessarily the same as the odds. Dismissing Robartes entirely, as Todhunter did, would be a mistake.

Robartes was very interested in probability problems. In the 1718 edition of the *Doctrine of Chances*, De Moivre quotes Robartes's work from the 1680s or 90s on the analysis of the dice game Raffle. Robartes also wrote a manuscript on two probability problems that was presented to the Royal Society on March 6, 1692 but never published (see Bellhouse 2011: p. 67). What is evident from this manuscript is that Robartes is using the binomial expansion to solve some problems in probability. This approach to probability is made explicit in the second volume of John Harris's (1710) scientific dictionary *Lexicon Technicum*. Under the heading "Play" there is a loose translation of Huygens's *De Ratiociniis*, along with some additional material. For the first few cases listed in *De Ratiociniis*, Harris gives Huygens's method for the division of stakes. For these cases, each solution is followed by another method that Harris calls the "method of unciæ of a binomial"; for all remaining cases Harris uses only the new method. He attributes this new method to Francis Robartes. It is essentially the binomial theorem applied to the problem of the division of stakes. The exact general algorithm that Robartes gives by way of Harris is repeated by De Moivre (1711) in his *De Mensura Sortis*.

4.4 Abraham De Moivre's Work in the Theory of Probability

De Moivre took a slightly different approach from that of Huygens to probability calculation by abandoning expected value calculations for the most part and focusing instead on

the enumeration of the number of outcomes favourable to an event and the number unfavourable. This approach is now known as the classical theory of probability. Admittedly, Montmort and Jakob Bernoulli both devoted enormous effort to developing combinatorial methods or methods of enumeration. It was, however, De Moivre who saw the classical theory of probability as a guiding principle in calculation. The classical theory became standard in Britain until the probability methodology of Laplace was introduced to Britain in the 1820s.

It was Robartes who encouraged De Moivre to work on problems in probability (see De Moivre 1711 and 1984: p. 237). Robartes had shown De Moivre a copy of Montmort's *Essay d'analyse sur les jeux de hazard*, which had been published in 1708. In De Moivre's words, Robartes "was pleased to propose to me some Problems of much greater difficulty than any he had found in that Book". The problems that Robartes gave De Moivre were two versions of a non-standard division of stakes problem in the context of two different situations in lawn bowling, as well as one more problem that is essentially the occupancy problem expressed in terms of finding the probability that specified faces on a generalized die show at least once in a number of throws of this generalized die. De Moivre claimed that he solved the first problem that Robartes had given him within a day of receiving it. The problem had been set as a challenge; Robartes had his own solution, but it was laborious. Given De Moivre's solution, Robartes posed two more problems and encouraged De Moivre to write on probability. The manuscript, *De Mensura Sortis*, was finished during a holiday that De Moivre took at a country house, possibly Robartes's. Since the resulting published paper was written in Latin, De Moivre probably had an international audience in mind.

While *De Mensura Sortis* was written in Latin, all editions of *Doctrine of Chances* were in English. Both Montmort and Johann Bernoulli questioned De Moivre's decision to write in English; the work would not be accessible to most mathematicians on the Continent. De Moivre did produce some papers on probability published in Latin in *Philosophical Transactions* (De Moivre 1714 and 1722); and there are probability results in his *Miscellanea Analytica* (De Moivre 1730). Further, he published in Latin at his own expense his normal approximation to the binomial (De Moivre 1733). Based on his comments in a 1717 letter to Brook Taylor (see Bellhouse 2011: pp. 116–17) and on his publications in Latin, De Moivre considered that he had made at least five major contributions to probability. These are: an approximation to the binomial, known eponymously as the Poisson approximation to the binomial; another approximation to the binomial, known as the normal approximation to the binomial; solutions to the duration of play problem, especially a trigonometric solution and another solution using recurring series; the solution to the problem of the pool, now known as Waldegrave's problem, for more than three players; and the introduction of the use of generating functions, initially to solve the problem of finding the probability of obtaining a given sum for the faces that show in the throw of several dice each with f faces. In every case, his solutions were original and brilliant. With one exception, modern historians of science would agree with De Moivre's self-assessment. The exception is Waldegrave's problem for four players. Hald (1990: p. 390) states that De Moivre's "method is theoretically simple but in practice very cumbersome for more than three players." This opinion was held earlier by De Moivre's contemporary and rival, Montmort (1713: pp. 361–70); but it did not seem to dampen De Moivre's pride in his result.

The duration of play problem is related to a challenge problem appearing in *De Ratiociniis*, known as the gambler's ruin problem. As translated by Edith Dudley Sylla, Huygens states the problem as (see Bernoulli 2006: p. 189):

> Each taking 12 coins, A and B play with three dice on the condition that if 11 points are thrown, A will give a coin to B, but if 14 points are thrown, B will give a coin to A. Whoever first gains all the coins wins. The ratio of the lot of A to the lot of B is found to be as 244,140,625 to 282,429,536,481.

The duration of play problem is to find the probability that one of the players wins all the coins within a certain number of plays, say m. In *De Mensura Sortis*, De Moivre considers the general situation of unequal skills or chances between the players and of the players having different capital. He provides a general solution using an algorithm based on a binomial expansion. Montmort (1713: pp. 361–70) criticized De Moivre's solution as being impractical when m is large – too many calculations are required. In response, De Moivre (1718: pp. 115–19) developed a method based on recurring series which he outlined in his *Doctrine of Chances*. The recurring series arises from the recurrence relation to be found in the duration of play problem. A player is ruined at the m^{th} round of play if he had only one unit of capital at the $(m-1)^{\text{th}}$ round and lost it at the next round. This player's probability of ruin at the m^{th} round of play is then the probability that he loses the m^{th} round multiplied by the probability that the game has continued to the $(m-1)^{\text{th}}$ round while leaving the player in question with one unit of capital. The sum of the ruin probabilities for each player for the first m games is the probability that the duration of play is at most m rounds of play. De Moivre (1718: pp. 149–50) also found a very elegant trigonometric solution to the duration of play problem that reduces the number of calculations tremendously. He gives hints of it in his 1718 *Doctrine of Chances* but held back publication of his proof until after Montmort's death (De Moivre 1722).

Waldegrave's problem is named for an Englishman, Francis Waldegrave, who posed the problem to Montmort while Waldegrave, a Jacobite, was exiled in France (see Bellhouse 2007). It arose from a real gambling situation in which a game that is typically played one-on-one is expanded to include more than two players. For three players, each one puts an equal ante into the pool or pot. Two of the players are chosen to play first; the other stays out. In every round of play the loser of the round puts an amount of money, always the same amount, into the pool. The winner plays the person who stayed out for the round. Play continues in like fashion until one of the players has beaten the other two successively. That player takes the pool. Early gambling literature held that it was advantageous to be the player to sit out first. By finding the chances of winning for each of the three players, De Moivre shows that it is actually disadvantageous to sit out first, provided that the amount the loser puts in at any round is less that 7/6 of each player's ante (De Moivre 1711 and 1984: pp. 249–51). Play can be expanded to include any number of players. To win the pool, a player must defeat all the others in succession.

Finding the number of chances to obtain a given sum on the faces that show in the throw of several dice was a standard problem for mathematicians. As noted already, Huygens had considered three dice in 1657 and Strode considered a few more dice in 1678. In the first edition of his *Essay d'analyse*, Montmort (1708) gave numerical results for up to nine dice. Unstated, but using mathematical induction and combinatorial arguments, De Moivre gives a general result in *De Mensura Sortis* for any number of dice, each with a given number

of sides. It was not until the 1720s that he discovered an elegant way to find the solution through a generating function; the method is explained in *Miscellanea Analytica* (De Moivre 1730: pp. 196–7). The idea was very simple. For example, if the binomial expansion $(a+b)^n$ is treated as an algebraic object, then the coefficient of the term $a^x b^{n-x}$ gives the number of chances of obtaining x "successes" and n–x "failures" in n independent Bernoulli trials where the probability of success for each trial is "fair", i.e. ½. This is the "method of the unciæ of a binomial" that both Robartes and De Moivre used to solve the division of stakes problem when the players have equal skills. For the sum of the faces that show in the throw of n dice, the generating function is a multinomial of the form $(a+a^2+a^3+\cdots a^f)^n$, where f is the number of faces on a die. The coefficient of the term a^x in the expansion is the number of chances to obtain x as the sum of the faces that show.

Later, in the second edition of the *Doctrine of Chances*, De Moivre (1738: p. 243) uses a generating function to solve the problem of runs. When r successes occur in a row in n independent Bernoulli trials, then there is a run of length r. De Moivre's generating function gives a simple solution to finding the probability of obtaining a run of length r.

De Moivre's first approximation to the binomial, the so-called Poisson approximation to the binomial, appears in *De Mensura Sortis* (De Moivre 1711 and 1984: pp. 239–40 and 242–3). It was obtained in the context of an extension to Huygens's and, unknown to De Moivre, Arbuthnot's previously mentioned solution to the dicing problem, which is to find the number of Bernoulli trials n necessary to obtain with probability p at least one success when the probability of success is $1/f$. De Moivre's extension to the problem is to set $p=\frac{1}{2}$ and to consider "at least two successes", "at least three successes", and so on. As part of his solution, De Moivre considers the situation in which n and f are infinitely large but $z=n/(f-1)$ is finite. The value of n to obtain, for example, at least four successes is found in the solution for z in the equation $z=\ln(2)+\ln\left(1+z+\frac{1}{2}z^2+\frac{1}{6}z^3\right)$. For a finite but large value of f then the solution for n can be found.

De Moivre began working on his second approximation to the binomial, now known as the normal approximation, in 1721 as a result of a problem suggested to him by Alexander Cuming, a Scottish baronet with mathematical interests. It is difficult to say what it was exactly that Cuming suggested to De Moivre. Most likely it was the following one, set in terms of a fictitious game of chance; Cuming and many others were fond of expressing probability problems in that way. Two players of equal skill to win, or, with equal probability of success, engage in n games. At the end of play, the player who has won the majority of games gives to a spectator an amount of money equal to the difference between the number of games he won and $\frac{n}{2}$. The problem is to find the expected amount that the spectator would receive. Since the solution is not obvious, it is likely that this was the problem posed by Cuming, and not any subsequent ones. The solution to the problem involves finding the middle term in the binomial expansion of $(1+1)^n = 1+\frac{n}{1}+\frac{n(n-1)}{2\times 1}+\cdots+\frac{n(n-1)\cdots 3\times 2\times 1}{n(n-1)\cdots 3\times 2\times 1}$. The associated binomial probability is the middle term in the expansion of $(1+1)^n$ divided by 2^n. This is given by

$$\frac{n!}{2^n\,(n/2)!\,(n/2)!}$$

when n is an even integer and where $n! = n(n-1)(n-2)\cdots 3\times 2\times 1$. The key to the normal approximation is to find an approximation to the term $n!$ when n is large. De Moivre

published his solution to Cuming's problem and his first stab at the approximation to the terms in the binomial expansion in his *Miscellanea Analytica* (De Moivre 1730: pp. 96–106). Spurred on by some insights and approximations to the logarithm of the factorial function provided by his friend James Stirling, De Moivre (1733) obtained his final and satisfactory approximation to the middle term of the binomial. This is given by

$$\frac{n!}{2^n\,(n/2)!\,(n/2)!} \cong \frac{\sqrt{2}}{\sqrt{n\pi}}$$

De Moivre's final step is to find approximations to other terms in the binomial based on his approximation to the middle term. Along the way he also considers unequal skill between the two players or unequal chances to win. He published his results in Latin at his own expense and circulated it to a number of mathematicians. See Stigler (1986: pp. 78–88) for a discussion of De Moivre's approximation to the binomial.

This new and powerful result probably motivated De Moivre to revise his *Doctrine of Chances*. He included an English translation of his normal approximation in the second edition of the *Doctrine of Chances*, published in 1738. A third edition was published posthumously in 1756. This final edition contains no new results of substance; there are additional explanations and examples that increased the size of the book by about one third over the second edition.

De Moivre's ability to carry out research in probability was always hampered by the time he needed to devote to making a living. For many years he worked as a tutor in mathematics to the sons of the aristocracy, wealthy merchants and other landed families. Later in life, he acted as a consultant on annuity valuations. These constraints on his time are manifest in the length of time that it took him to produce the first and second editions of the *Doctrine of Chances*. There is evidence that he began working on the first edition at least three years before it was published. Similarly, the key result that was included in the second edition, published in 1738, was obtained in De Moivre (1733). Advertisements for subscribers to the second edition as well as to his *Miscellanea Analytica* appeared two years or more before each book was finally published (see Bellhouse 2011: pp. 173–4 and 188).

De Mensura Sortis and especially all editions of the *Doctrine of Chances* contain a number of analyses of card and dice games, some suggested by colleagues and friends of De Moivre. Unlike Arbuthnot's *Of the Laws of Chance*, *De Mensura Sortis* and the three editions of *Doctrine of Chances* were not gambling manuals and they were not for the faint of heart. Written in the same spirit as *De Ratiociniis in Ludo Aleae*, De Moivre's books on probablity, beginning with *De Mensura Sortis*, contain solutions to very difficult mathematics problems set in the context of games of chance. This is evident in two comments made about *De Mensura Sortis*. The first is by William Browne, who tried to explain why he had bothered to translate Huygens's *De Ratiociniis* after De Moivre had produced his work. He states (Huygens 1714):

> M. De Moivre's Piece therefore will be very far from lessening the Worth of Mr. Huygens's; and the superficial Mathematicians will still be glad to satisfy their Enquiries by this last Author's easy, tho' more tedious Method, as not being able to understand the other's more comprehensive and general one; while those of a greater Depth, will with no less delight first read M. Huygens's Treatise, in order to proceed with so much greater Pleasure afterwards to peruse M. De Moivre's Additions …

The second is from a reviewer of *De Mensura Sortis* who states in his 1712 review that the work would be applauded in the halls of the academy, but would be of little use in the halls of gaming (see Bellhouse 2011: p. 70).

4.5 Life Annuities: The Eighteenth-century Application of Probability

Although later in life De Moivre acted as a mathematical consultant to gamesters on games of chance, for the most part his work in probability was motivated by finding solutions to challenging mathematical problems. There is one glaring exception. De Moivre (1725) carried out some fundamental work on life annuities. This work was motivated by his connection to one of his aristocratic patrons, Thomas Parker, 1^{st} Earl of Macclesfield. More than a decade earlier, Nicolas Bernoulli had encouraged De Moivre to work on applications of probability to economics and politics. De Moivre had put aside Bernoulli's suggestion, claiming he was too busy with his teaching (Bellhouse 2011: p. 153).

Bernoulli's suggestion was general and vague. Judging from what De Moivre produced in his *Annuities Upon Lives*, Macclesfield's questions that motivated De Moivre were probably much more specific and related to land. Macclesfield, a newly minted aristocrat, was land-acquisitive. As chief judge in the High Court of Chancery, he handled a number of cases related to the value of property. Many of De Moivre's life annuity calculations are connected to property through a method of land leasing called "leases for lives". In one type of lease, which has its roots as far back as the reign of Henry VIII, the length of the lease ran until the last survivor of three people named on the lease died. The value of future rents would then be the present value of a last-survivor annuity on three lives. If one of the persons named on the lease died, the lease could be renewed by naming a replacement and paying a fee for this privilege. The valuation of the fee is another exercise in applied probability based on life contingencies. From an expectant landlord's point of view there are other valuations of interest. A landlord might hold his land on the condition that on his death the lease-holding passes to someone else, the expectant landlord who holds what is called 'the reversion'. That expectant landlord might want to sell his reversion well before the original landlord's death. This requires a valuation of the reversion. De Moivre tackled and provided solutions to all these, and more, life-contingent questions. Inspired by Edmond Halley's life table, published in 1693, De Moivre modeled the survivor curve as a linear function of age. Based on this assumption, he could get simple forms for producing the values of life annuities that were functions of the values of annuities certain. This cut down enormously on the amount of calculation required to make a valuation. See Bellhouse (2011: pp. 153–71) for a full discussion.

De Moivre considered his work on life annuities as part of his overall program in probability. All of his results in this area were included in both the second and the third editions of the *Doctrine of Chances* (De Moivre 1738: pp. 211–31 and De Moivre 1756: pp. 261–328).

None of the work on annuities was translated into Latin, so De Moivre must have considered the topic "of local interest" only.

A very able mathematician and expositor of mathematics, Thomas Simpson was both an interpreter of De Moivre's work and a competitor with De Moivre for new results. Simpson's (1740) *The Nature and Laws of Chance* is closely related to the second edition of *Doctrine of Chances*, published in 1738. Most of De Moivre's problems in *Doctrine of Chances* appear in Simpson's book. Where De Moivre has not provided an explicit proof, Simpson has filled the gap. Where De Moivre has provided a proof, Simpson has provided the same proof or an alternate and sometimes simpler proof. Simpson's (1742) *Doctrine of Annuities and Reversions* contains accounts of the same problems in life contingencies as appear in De Moivre's *Annuities Upon Lives*. The major difference is that Simpson's calculations and approximations are based on new life tables constructed for the City of London, while De Moivre's calculations are based on his survivor function, modeled from Halley's life table that was constructed for the City of Breslau. Both of Simpson's books were priced lower than De Moivre's, which led to some acerbic comments from De Moivre.

4.6 Thomas Bayes

De Moivre had a student whom he tutored in mathematics, a pious aristocrat named Philip Stanhope, 2nd Earl of Stanhope. On more than one occasion this student visited Tunbridge Wells, about fifteen miles from his country estate. There he met and became friends with an obscure Presbyterian minister named Thomas Bayes, whom Stanhope later sponsored for fellowship in the Royal Society. They corresponded on a number of mathematical subjects, including probability. This is the most likely route that led Thomas Bayes to taking such an interest in probability (Bellhouse 2011: p. 201).

Bayes made two contributions to probability, one minor and the other major, but both generally ignored even into the nineteenth century. Both contributions appeared posthumously in papers published in *Philosophical Transactions*, and were probably inspired by De Moivre's work in probability, especially the normal approximation to the binomial. Both papers were written in the form of letters to John Canton, a friend of Bayes and, like Bayes, a fellow of the Royal Society and dissenter in religious matters. The minor contribution was probably sent to Canton by Bayes in the 1740s. The major contribution was found by Richard Price among Bayes's papers after his death. Price, another of Bayes's friends and fellow Presbyterian minister, wrote the letter to Canton that contained Bayes's contribution. For a biography of Bayes, see Bellhouse (2004).

Bayes's minor contribution is directly related to De Moivre's normal approximation to the binomial, which depends on an approximation to the natural logarithm of the factorial function. Both De Moivre and James Stirling had obtained approximations to the factorial function. Bayes pointed out that these approximations were based on infinite series that did not converge. He went on to provide a convergent series and hence fixed a minor glitch in the derivation of the normal approximation to the binomial. Appearing in *Philosophical Transactions*, the paper (Bayes 1763a) was read before the Royal Society on November 24, 1763. Bayes provided the same paper to Philip Stanhope; it was dated 1747 (Bellhouse 2002).

In eighteenth-century eyes, Bayes's major contribution to probability was also intimately connected to De Moivre's normal approximation to the binomial. De Moivre's approximation was interpreted as a mathematical law describing the stability of the proportion of successes to be observed in a long run of observations of successes and failures. Bayes's (1763b) result was in the other direction: given an observed proportion of successes from the past, one can find the probability that the next observation lies within a given interval of values. Bayes's model used a square flat table, akin to a billiard table without pockets, in which a ball was rolled across the table to determine a line parallel to two opposite sides of the table. The line marked a demarcation between success to one side and failure to the other. Then a number of balls were rolled across the table and the proportion of successes noted. When a new ball was about to be rolled across the table, the probability that the ball would stop between two given lines was sought. The numerical calculation of the probability based on a uniform prior generated by the square table model requires the evaluation of incomplete beta functions. Bayes used infinite series approximations to do this. In most cases, a large amount of calculation is involved in the series approximation (see Stigler 1986: p. 130). Later, Price (1764) tried to find an improved approximation but was no more successful than Bayes.

Although Bayes's (1763b) insight is fundamental to inference about future outcomes based on past experience, the value of this insight was not recognized by his fellow countrymen. Their focus instead was on the onerous amount of calculation. In a history of the Royal Society, Thomson (1812: p. 306) describes Bayes's paper and concludes, "The solution is much too long and intricate to be of much practical utility." It would not be until De Morgan introduced Laplace's work in probability to Britain in the 1830s, work that contained Laplace's version of Bayes's Theorem (which is the modern simple version of it found in introductory textbooks), that the Bayesian approach to inference achieved any traction in Britain.

What motivated Bayes to work on his billiard-inspired problem is unknown. None of Bayes's known papers give any hint as to why he considered any of his mathematical problems. In his manuscripts, mathematical problems are stated and then solved; nothing more is given (Bellhouse 2002). On the other hand, Richard Price's motivation in bringing Bayes's work to print was distinctly theological and philosophical. To Price, Bayes's probability result was an argument for the existence of God. De Moivre's central limit theorem had demonstrated a mathematical regularity under which things happened and did not happen over the long run. Bayes's result was a stronger argument in that it showed, in events that could occur again and again, that "recurrency or order is derived from stable causes or regulations in nature, and not from any of the irregularities of chance" (Bayes 1763b: p. 374). Further, based on the title of the offprints of Bayes's paper commissioned by Price, Stigler (2013) has argued convincingly that Bayes's paper was written in response to David Hume's essay on the theological question of miracles.

4.7 Discussion

Stigler (1988) describes the latter half of the seventeenth century as the Dark Ages of probability in which, on the surface, little happened in the subject. Hald (1990) calls the

decade of 1708–18 the great leap forward in probability. The development of probability in Britain falls into line with these categories. Prior to 1711 there was very little development of the mathematical theory of probability in England – only minor contributions by Newton, Robartes, and Strode. More effort was made at the application of probability to areas such as gambling, politics, medicine, and theology. The work was scattered and unconnected, other than in that most of it was influenced by the work of Christiaan Huygens. De Moivre's *De Mensura Sortis*, one of the key publications in Hald's great leap forward, was pivotal in the development of probability and its applications in England. Through his initial publication and the editions of his *Doctrine of Chances*, De Moivre solved most of the difficult problems of his day in probability. The major application of probability, the valuation of life annuities, was also pioneered by De Moivre; and the impact of his work in this area was felt well into the nineteenth century. A major application today of probability theory, Bayes's theorem, was espoused within a decade of De Moivre's death in 1754, but not embraced until well into the next century.

No other major advances in probability were made in Britain during the eighteenth century. New developments had to wait until the third and fourth decades of the nineteenth century. At that time there were two very separate and different stimuli. On the academic side, Augustus De Morgan, for example, introduced the work of Laplace into Britain, including the modern version of Bayes's theorem, and carried out his own research in probability. On the economic side, the rising middle class, a result of the Industrial Revolution, created a demand for life insurance that spurred new developments in the actuarial science that De Moivre, and Halley before him, had engendered.

References

Arbuthnot, J. (1692) *Of the Laws of Chance, or, a Method of Calculating the Hazards of Game*. London: Motte.

Arbuthnott, J. (1710) An argument for Divine Providence, taken from the constant regularity observ'd in the births of both sexes. *Philosophical Transactions*. 27. pp. 186–90.

Bayes, T. (1763a) A letter from the late Reverend Mr. Thomas Bayes to John Canton, M.A. & F.R.S. *Philosophical Transactions of the Royal Society*. 53. pp. 269–71.

Bayes, T. (1763b) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society*. 53. pp. 370–418.

Bellhouse, D. R. (1989) A manuscript on chance written by John Arbuthnot. *International Statistical Review*. 57. pp. 249–59.

Bellhouse, D. R. (1993) The role of roguery in the history of probability. *Statistical Science*. 8. pp. 410–30.

Bellhouse, D. R. (2000) *De Vetula*: a medieval manuscript containing probability calculations. *International Statistical Review*. 68. pp. 123–36.

Bellhouse, D. R. (2002) On some recently discovered manuscripts of Thomas Bayes. *Historia Mathematica*. 29. pp. 383–94.

Bellhouse, D. R. (2004) The Reverend Thomas Bayes, FRS: a biography to celebrate the tercentenary of his birth (with discussion). *Statistical Science*. 19. pp. 3–43.

Bellhouse, D. R. (2007) The problem of Waldegrave. *Journal Electronique d'Histoire des Probabilités et de la Statistique*. 3 (2). [Online] Available from: http://www.jehps.net/decembre2007.html. [Accessed 25 Aug 2015.]

Bellhouse, D. R. (2011) *Abraham De Moivre*: *Setting the Stage for Classical Probability and Its Applications*. Natick, MA: CRC Press.

Bellhouse, D. R. and Genest, C. (2007) Maty's life of Abraham De Moivre, translated, annotated and augmented. *Statistical Science*. 22. pp. 109–36.

Bernoulli, J. (2006) *The Art of Conjecturing*: *Together with a Letter to a Friend on Sets in Court Tennis*. Translated from the Latin by E. M. Sylla. Baltimore, MD: Johns Hopkins Press.

Cumberland, R. (2005) *A Treatise of the Laws of Nature*. Translated from the Latin by J. Maxwell. Indianapolis, IN: Liberty Fund.

De Moivre, A. (1711) De mensura sortis seu; de probabilitate eventum in ludis a casu fortuito pendentibus. *Philosophical Transactions*. 27. pp. 213–64.

De Moivre, A. (1714) Solutio generalis altera praecedentis, ope combinationum & serierum infinitarum. *Philosophical Transactions*. 29. pp. 145–58.

De Moivre, A. (1718) *The Doctrine of Chances, or, a Method of Calculating the Probability of Events in Play*. London: W. Pearson.

De Moivre, A. (1722) De fractionibus algebraicis radicalitate immunibus ad fractiones simpliciores reducendis, deque summandis terminis quarumdam serierum aequali intervallo a se distantibus. *Philosophical Transactions of the Royal Society*. 32. pp. 162–78.

De Moivre, A. (1725) *Annuities Upon Lives, or, The valuation of Annuities upon any Number of Lives, as also, of Reversions to which is added, an Appendix Concerning the Expectations of Life, and Probabilities of Survivorship*. London: W. Pearson.

De Moivre, A. (1730) *Miscellanea Analytica de Seriebus et Quadraturis*. London: J. Tonson and J. Watts.

De Moivre, A. (1733) Approximatio ad Summam Terminorum Binomii $\overline{a+b}|^n$ in Seriem expansi. Self-published paper.

De Moivre, A. (1738) *The Doctrine of Chances: or, a Method of Calculating the Probability of the Events in Play*. 2nd ed. London: Woodfall.

De Moivre, A. (1756) *The Doctrine of Chances: Or, a Method of Calculating the Probability of Events in Play*. 3rd ed. London: Millar.

De Moivre, A. (1984) On the measurement of chance, or, on the probability of events in games depending on fortuitous chance. *International Statistical Review*. 52. pp. 237–62. Translated from the Latin by Bruce McClintock.

Friesen, J. (2003) Archibald Pitcairne, David Gregory and the Scottish origins of English Tory Newtonianism. *History of Science*. 41. pp. 163–91.

Guerrini, A. (1986) The Tory Newtonians: Gregory, Pitcairne, and their circle. *The Journal of British Studies*. 25. pp. 288–311.

Hald, A. (1990) *A History of Probability and Statistics and Their Applications before 1750*. New York, NY: John Wiley.

Harris, J. (1710) *Lexicon technicum: or, an Universal English Dictionary of Arts and Sciences: Explaining Not only the terms of Art, but the Arts themselves.* Vol. II. London: Brown et al..

Harvey, W. (1628) *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus*. Frankfurt: Fitzer.

Hobbes, T. (1994) *Leviathan with Selected Variants from the Latin Edition of 1668*. Cambridge, MA: Hackett Publishing Company.

Huygens, C. (1657) De Ratiociniis in Ludo Aleae. In van Schooten, F. (ed.), *Exercitationum Mathematicarum libri quinque*. pp. 517–34. Leiden: Elsevier.

Huygens, C. (1714) *Christiani Hugenii libellus de ratiociniis in ludo aleæ. Or, the value of all chances in games of fortune; Cards, Dice, Wagers, Lotteries, &c.*. Translated from the Latin by William Browne. London: Keimer and Woodward.

Kahle, L. M. (1735) *Elementa Logicae Probabilium Methodo Mathematica in Usum Scientarum et Vitae Adornata*. Braunschweig: Renger.

Maty, M. (1755) Mémoire sur la vie & sur les écrits de Mr. de Moivre. *Journal britannique*. 18. pp. 1–51.

Montmort, P. R. de (1708) *Essay d'analyse sur les jeux de hazard*. Paris: Quillau.

Montmort, P. R. de (1713) *Essay d'analyse sur les jeux de hazard*. 2nd ed. Paris: Quillau.

Nash, R. (1991) *John Craige's Mathematical Principles of Christian Theology*. Carbondale, IL: Southern Illinois University Press.

Newton, I. (1967) *The Mathematical Papers of Isaac Newton*. Vol. 1 1664–1666. Cambridge: Cambridge University Press.

Parkin, J. (1999) *Science, Religion and Politics in Restoration England*: *Richard Cumberland's De Legibus Naturae*. Woodbridge: Boydell Press.

Pepys, S. (1926) *Private Correspondence and Miscellaneous Papers of Samuel Pepys 1679–1703 in the Possession of J. Pepys Cockerell*. Gravesend: Bell.

Pitcairne, A. (1693) *Dissertatio de Motu Sanguinis per Vasa Minima*. Leiden: Elzevier.

Price, R. (1764) A demonstration of the second rule in the essay towards the solution of a problem in the doctrine of chances, published in the Philosophical Transactions, Vol. LIII. Communicated by the Rev. Mr. Richard Price, in a Letter to Mr. John Canton, M.A. F.R.S.. *Philosophical Transactions*. 54. pp. 296–325.

Roberts, F. (1693) An arithmetical paradox, concerning the chances of lotteries. *Philosophical Transactions*. 17. pp. 677–81.

Simpson, T. (1740) *The Nature and Laws of Chance*. London: Edward Cave.

Simpson, T. (1742) *The Doctrine of Annuities and Reversions, Deduced from General and Evident Principles*. London: J. Nourse.

Stigler, S. M. (1986) *The History of Statistics*: *The Measurement of Uncertainty before 1900*. Cambridge, MA: Belknap Press.

Stigler, S. M. (1988) The Dark Ages of probability in England: the work of Richard Cumberland and Thomas Strode. *International Statistical Review*. 56. pp. 75–88.

Stigler, S. M. (1999) *Statistics on the Table*: *The History of Statistical Concepts and Methods*. Cambridge, MA: Harvard University Press.

Stigler, S. M. (2006) Isaac Newton as a Probabilist. *Statistical Science*. 21. pp. 400–03.

Stigler, S. M. (2013) The true title of Bayes's essay. *Statistical Science*. 28. pp. 283–8.

Strode, T. (1678) *A Short Treatise of the Combinations*, *Elections*, *Permutations & Composition of Quantities*. London: W. Godbid and Wyer.

Thomson, T. (1812) *History of the Royal Society, from its Institution to the End of the Eighteenth Century*. London: Robert Baldwin.

Todhunter, I. (1865) *A History of the Mathematical Theory of Probability from the Time of Pascal to that of Laplace*. Cambridge: Cambridge University Press.

CHAPTER 5

A BRIEF HISTORY OF PROBABILITY THEORY FROM 1810 TO 1940

HANS FISCHER

5.1 Introduction

As an independent discipline of mathematics, probability theory was essentially shaped during the first decades of the 20th century. In the 18th and 19th centuries, probability was conceived rather as a part of moral and natural sciences, and not of mathematics proper. Correspondingly, "applied" fields, such as probabilities of testimonies or the correctness of human decisions, the theory of observational errors, the kinetic theory of gases, and topics which, from today's point of view, would be included within the field of statistical inference, were in the focus of probability theory and defined its status within the whole complex of mathematical sciences until the last third of the 19th century. Still, after the publication of Laplace's 1812 *Théorie analytique des probabilités*, which served as a paradigm of probabilistic research for most of the 19th century, specific analytic methods of probability increasingly aroused the interest of mathematicians, and thus particular problems of probability likewise began to develop a purely mathematical quality. Applied fields, such as statistical physics or financial mathematics, maintained a decisive and enduring influence upon the development of probability theory, but in the 20th century the mathematical essence of the theory attained complete autonomy, and thus gradually came to constitute "modern" probability. Significant steps in this development were the study of limit theorems, in particular the central limit theorem, without considering any extra-mathematical contents, and the discussion of infinite events which cannot be conceived as an obvious continuation of series of real random experiments. With the introduction of a measure-theoretic framework for such investigations, an additional and crucial element fostering the growing abstraction of probabilistic subjects was established. In this way, the main subfields of modern probability, such as axiomatics, weak and strong limit theorems, sequences of non-independent random variables, and stochastic

processes, which had originally emerged in a rather disparate way, could be integrated into a well-connected complex of modern probability theory until the beginning of World War II.

In this chapter, the historical development outlined above will be explained in fair detail, but, because space is limited in the present volume, full information cannot be given on all relevant primary sources. Summarizing works which provide comprehensive bibliographies are referred to in the text at several places, however. In particular, see Hald (1998) and von Plato (1994) for 19th- and 20th-century sources, respectively.

5.2 Laplacian Probability

Laplace's first significant probabilistic contribution, published in 1774, already shows his specific interest in asymptotic problems. Laplace's concern was to determine the true ratio of black and white balls in an urn containing "infinitely many" such balls, if, after n drawings with replacement, p black and $q = n - p$ white balls had occurred. In a way very similar to Thomas Bayes's approach—we can only speculate about whether Laplace knew his 1764 contribution—Laplace calculated the conditional probability that the ratio r between the true number of all black balls and all balls in the urn is between $p/n - \omega$ and $p/n + \omega$ (the positive quantity ω possibly depending on n) if the relative frequency $h_n = p/n$ of black balls has been observed. In so doing he assumed a uniform a priori distribution for all possible values of r between 0 and 1. To the result he applied his later so-called "Laplacian approximation method" for approximating integrals depending on very large parameters, and in this way he arrived at

$$\lim_{n\to\infty} P(|r - h_n| \leq \varepsilon(n) \,|h_n) \to 1 \left(\varepsilon(n) = \mathrm{O}\left(n^{-\alpha}\right), \frac{1}{3} < \alpha < \frac{1}{2}\right) \tag{5.1}$$

(expressed in modern terms). In a similar way, a good deal of Laplace's work concerned inverse probabilities, i.e., conditional "a posteriori" probabilities $P(H|R)$ for certain "hypotheses" H, calculated according to Bayes's formula after the observation of statistical data R. In most cases the tacit assumption of an a priori equiprobability of all possible hypotheses was made. This field of research included problems such as the probability of a boy's birth, the comparison of birth rates of boys at several places in Europe, and similar problems in population statistics. Influenced by Condorcet's approach, Laplace also tried to determine probabilities for the correctness of human decisions by Bayesian methods. Laplace was cautious about oversimplified assumptions, but he also stressed the usefulness of probabilistic considerations in this context. His most significant result concerned the probability that a defendant is actually guilty if there is a certain ratio of votes against him.

One issue with which Laplace was occupied from his earliest work in probability concerned sums of independent random variables, e.g., in relation to arithmetic means of angles of inclination of comets' orbits. Around 1810, by applying characteristic functions (see (5.7) below) in a rudimentary form, Laplace found approximate formulae for the respective probabilities. His results correspond to local (i.e., assertions on the convergence of densities or single discrete probabilities) and integral (i.e., assertions on probabilities assigned to a whole interval) central limit theorems. In the following, "central limit

theorem" will be abbreviated as "CLT." Laplace's CLTs referred to "large" numbers of independent random variables $X_1, \ldots, X_n$, where, with the exception only of the particular case of two-valued discrete variables, all of them were assumed to have the same density concentrated on a compact interval. In modern terms, Laplace's integral CLT can, in the case of identically distributed random variables, be expressed by

$$P\left(\left|\sum_{k=1}^{n}\lambda_k(X_k-\mu)\right| \le a\sqrt{\sum_{k=1}^{n}\lambda_k^2}\right) \approx \frac{2}{\sigma\sqrt{2\pi}}\int_0^a e^{-\frac{x^2}{2\sigma^2}}\,dx,$$

where λ_k designate arbitrary real multipliers, and $\mu := \mathrm{E}X_1, \sigma^2 := \mathrm{Var}X_1$. Laplace's CLTs were at no place stated in a general form, but in each case derived anew with relation to their specific applications. These included sums or arithmetic means of various stochastically fluctuating quantities in nature and society, as, e.g., mean air pressure, mean duration of life, gains in gambling, or an insurance company's profit. In this context, statements along the lines of today's weak laws of large numbers were established, and tests of hypotheses on hidden regularities could be accomplished where the test statistic is the sum or arithmetic mean of a great many independent random variables. In this way, Laplace also showed that there was no reason for assuming the existence of a *cause primitive* (as he put it in French) influencing the particular position of comets' orbits.

A very prominent area of application of the CLT was error theory, which referred mainly to the probabilistic discussion of observational data in astronomy and geodesy under the assumption of all measurements being affected by errors that were conceived as random variables. Presumably, the occurrence of the normal distribution in Gauss's first argument (expounded in an 1809 paper) in favor of the method of least squares as applied in the framework of linear models related to astronomical observations, inspired Laplace to a new—asymptotic—justification for the method of least squares by aid of a CLT for linear combinations of observational errors.

Laplace's *Théorie analytique des probabilités*, henceforth abbreviated as *TAP*, of which the first edition appeared in 1812, was mainly a compilation of his thitherto achieved analytic and probabilistic results. The first part of the book contained an exposition of his most preferred analytical problems and methods, such as difference equations and generating functions, evaluation of particular definite integrals, or approximations of definite integrals whose integrands depend on large parameters. His pertinent results had a significant influence on the subsequent development of analysis, e.g., on complex integration or asymptotic expansions. The second, probabilistic, part of the *TAP* focused mainly upon statistics via Bayesian methods and problems surrounding the CLT, including error theory.

From the second edition (1814) of the *TAP* onward, a comprehensive introduction, the "*Essai philosophique sur les probabilities*," was added to the book, which later also appeared in many separate editions and different languages. The "*Essai*" not only disseminated the message of the universal applicability of probability theory to all physical, social, and moral problems, culminating in the famous saying: "Probability theory is basically good sense reduced to a calculus," but also dealt exhaustively with philosophical concepts of probability. Laplace based his considerations on a strict determinism, asserting that all processes in the world could be precalculated according to distinct dynamical laws, at least by a superhuman being, the proverbial "demon." Man, however, with only a restricted knowledge of causes and laws, was provided with the powerful tool of probability calculus. According to this idea,

the subjective conception of probability was emphasized by Laplace. At numerous places, however, he also hinted at the frequentist notion of probability.

Up to the end of the 19th century, there were hardly any accounts of probability theory that provided basically new stochastic concepts beyond those found in Laplace's work. Owing to the difficulties of the *TAP* regarding analytical methods and style, only a few mathematicians read this, as De Morgan put it, "Mont Blanc of mathematics," persistently and carefully. Thus, the impact of *TAP* was rather indirect, but Laplace's analytic theory, in particular in relation to asymptotic considerations on sums of independent random variables, motivated ambitious mathematicians to further research. Some basic ideas expounded in the "*Essai*," such as the universal applicability of probability theory, or the equivalence, at least in principle, of social and physical processes, had an enduring influence on the development of statistics.

5.3 Gaussian Error Theory

Carl Friedrich Gauss's contributions to probability were, practically without exception, in the framework of error theory. Especially since his two justifications of the method of least squares (in 1809 and 1823) appeared to be easier to understand than Laplace's asymptotic considerations, Gaussian error theory became rather popular. In his first justification, Gauss basically showed that if one assumes that the arithmetic mean is the maximum likelihood estimator in the case of direct observations, this implies a normal distribution as the law of error. From this, the method of least squares can be deduced as leading to the "most probable" estimations. As a consequence of Hagen's and Bessel's work, the rather contrived reasoning concerning the arithmetic mean was in many instances replaced by later authors with a direct derivation of the normal distribution as error law on the basis of the so-called "hypothesis of elementary errors:" Each error of observation proper was considered to be additively composed from a great many small elementary errors acting independently of each other. In this way the hypothesis of elementary errors became an important application of the CLT, and, in turn, also stimulated its further advancement. In his second justification Gauss showed that in linear models as applied to astronomical or geodetic observations, the variance of the deviation between an estimator (gained by a linear combination of the observations) and the respective true value is a minimum if the estimator is constructed according to the method of least squares.

In the framework of error theory, Gauss contributed much to the theory of probability, and thus significantly influenced its development. Many general notions and relations with respect to random variables are due to Gauss: the significance of variance (*mittlerer quadratischer Fehler*), and the fact that the variance of a sum of independent random variables is equal to the sum of the variances of the individual random variables, to give only an elementary example. Gauss derived an inequality for probabilities of a type similar to that of Bienaymé and Chebyshev; his discussion of higher moments (i.e., the expected values of natural powers) of observational errors laid the groundwork for later moment methods.

5.4 Poisson: Central Limit Theorem, Law of Large Numbers, Probabilities of Judgements

Regarding probability theory, Siméon Denis Poisson can be considered the most important follower of Laplace. In the field of limit theorems, Poisson elaborated and refined Laplace's analytic methods, and he generalized the assertion of the CLT. By aid of the CLT he established "his" law of large numbers, which would become the leitmotif of the probability of mass phenomena during the second half of the 19th century. In the field of probabilistic treatment of decisions in court trials he extended Laplace's approach by an ingenious inclusion of statistical data. His pertinent work was heavily criticized from the philosophical point of view, however; Poisson's account was considered an abuse of mathematics with respect to the free will of man. In this way, Poisson unwillingly contributed to the decline of the probabilistic discussion of moral problems, and altogether to the decline of Bayesian methods, from ca. 1840 on.

Thus, Poisson's major probabilistic work, *Recherches sur la probabilité des jugements en matière criminelle et en matière civile, précédées des règles générales du calcul des probabilités*, which was issued in 1837, had significant influence not by the topic which was emphasized in the title and to which the second part of the book was devoted, but through its first, "general" part. Poisson stated and proved the CLT (still in the sense of an approximate assertion for a large number of random variables) for general "quantities," which were essentially assumed to be uniformly bounded but differently distributed, and in this way he established an early intuitive concept of random variable. He also discussed, in terms of characteristic functions, (not entirely complete) conditions for the CLT as well as counterexamples to the CLT. The most important counterexample concerned sums of identically distributed independent random variables, each obeying a distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)},$$

the later (erroneously!) so-called "Cauchy distribution." Poisson's ideas were influential upon many later proofs of the CLT, including Lyapunov's around 1900.

Poisson's "law of large numbers" is not to be confused with the modern notion of this term. It was rather conceived as a sort of natural law asserting the stability of relative frequencies (or more generally of arithmetic means) among different series of observations in various situations. Poisson tried to explain this law by a two-stage model of causation. In this context, he derived from the CLT an auxiliary theorem which corresponds to the modern version of the weak law of large numbers for arithmetic means of independent random variables X_k:

$$\forall \varepsilon > 0 : P\left(\left|\frac{\sum_{k=1}^{n}(X_k - \mathrm{E}X_k)}{n}\right| > \varepsilon\right) \to 0 \quad (n \to \infty). \tag{5.2}$$

5.5 Philosophy of Probability in the 19th Century

Until the early 19th century, the now common distinction between "objective" and "subjective" probabilities was almost never made by mathematicians or philosophers, including Laplace. According to the paradigm of association of ideas, a "sound mind" was able to continuously register impressions in an unbiased way and thus currently modify his own subjective probabilities to a limit of objective character (see Daston 1988, chapter 4). This process corresponded to Laplace's 1774 result (5.1) on the increasing proximity of the true success probability r and the observed relative frequencies h_n.

After the Enlightenment, the basic idea of "good sense" disintegrated, and thus the close relation between subjective and objective probability was likewise no longer considered a matter of course. In his 1842 monograph "*Versuch einer Kritik der Principien der Wahrscheinlichkeitsrechnung*," Jakob Friedrich Fries combined a strictly deterministic world view with a frequentist interpretation of numerical probabilities which could be gained only by logical considerations regarding the symmetry of a perfect cube, for example, or by average ratios among series of observations, presupposing the stability of underlying causes. On the other hand, subjective probabilities could not be quantified in a sufficiently precise way, and therefore were not usable as a basis for mathematical considerations, according to Fries. A prominent proponent of objective probability together with a world view that was no longer completely deterministic was Augustin Cournot. Dispensing with Laplace's universal mechanistic conception, Cournot, alluding to Laplace's demon, argued in his 1843 book *Exposition de la théorie des chances et des probabilités* that even a supernatural being was not able to precisely precalculate all processes in nature and society. Therefore, randomness was not entirely due to human ignorance but had objective aspects also. Cournot anticipated ideas which, some 60 years later, were taken up anew in the context of discussions of determinism in dynamical systems, e.g., by Henri Poincaré.

Apparently influenced by Fries's exposition as well as the contemporary development of statistical physics, Johannes von Kries gave a very precise (but also very wordy) account of philosophy of probability in his 1886 book *Die Principien der Wahrscheinlichkeits-Rechnung*. Von Kries introduced the concept of *Spielraum*, which in some respect (and with some hindsight) can be compared with a sample space generated by the union of countably many irreducible and equiprobable events. In a way similar to those of Fries and Cournot, von Kries also discussed "philosophical" probabilities connected with rational guesses, which, however, had a semi-quantitative character at best. In contrast to those philosophical probabilities, "numerical" probabilities could be obtained only by a careful "logical" consideration leading to an appropriate *Spielraum*, which, however, could not be established on an entirely objective basis, except in singular cases. In 1889 Joseph Bertrand addressed similar ideas in context with his discussion of different probabilities assigned to the same random experiment as dependent on a closer definition of the random events under consideration. Von Kries also reflected upon "empiric" probabilities, which could be gained from relative frequencies, but only if the stability of the pertinent average numbers could be assumed. As a means for investigating stability, Fries favored Wilhelm Lexis's statistical theory of dispersion. Von Kries had a considerable influence on the foundations of probability theory right up to the first decades of the 20th century, at least in German-speaking Europe.

Together with the vanishing emphasis on subjective probability, the investigation of individual behavior (e.g., in human decisions) experienced a considerable loss of significance. In accord with the objective conception of probability, only a few topics remained in the foreground of contemporary work on probability, in particular with respect to "large numbers."

5.6 Further Discussions of Laplacian Error Theory

The chief problem of Laplacian error theory was the investigation of the asymptotic behavior of the distribution of

$$\sum_{k=0}^{n} \lambda_k \epsilon_k,$$

where λ_k designated real multipliers, and ϵ_k stochastically independent errors of observation with densities. A very notable contribution was due to Cauchy. In his scientific controversy with Bienaymé on the supremacy of the method of least squares during the summer months of the year 1853, he showed that

$$\left| P\left(-\nu \leq \sum_{j=1}^{n} \lambda_j \epsilon_j \leq \nu \right) - \frac{2}{\sqrt{\pi}} \int_0^{\frac{\nu}{2\sqrt{c\Lambda}}} e^{-\theta^2} d\theta \right| \leq C(n,\nu),$$

such that $C(n,\nu)$ (whose analytic expression was explicitly determined) tends to 0 for $n \to \infty$. All errors were presupposed to have the same symmetric and sufficiently smooth density f concentrated on a compact support $[-\kappa, \kappa]$, with $c := \int_0^{\kappa} x^2 f(x)\, dx$; the multipliers λ_j were supposed to be essentially of the order $\frac{1}{n}$, such that $\Lambda := \sum_{j=1}^{n} \lambda_j^2$ was of this order as well. In this way, Cauchy gave—if in a sketchy form only—arguments for a rigorous proof of an assertion equivalent to a fairly general CLT, which are sound even from the point of view of modern analysis. Already a few years earlier than Cauchy, Dirichlet had expounded similar ideas in one of his lecture courses in 1846, the notes of which remained unpublished (see Fischer 2011, section 2.4).

The above-mentioned Cauchy-Bienaymé controversy produced a second highlight of 19th-century probability: In order to give a plausibility argument in favor of least squares, Bienaymé (1853) deduced the relation

$$P\left(\left| \sum_{j=1}^{n} \lambda_j \epsilon_j \right| \leq t\sqrt{2\sigma^2} \right) = 1 - \frac{\theta f}{2t^2} \sum_{j=1}^{n} \lambda_j^2$$

for independent identically and discretely distributed errors ϵ_j with variance σ^2 and zero mean. The symbols θ and f denote positive quantities less than 1, depending on the law of error and on the multipliers λ_j. Bienaymé's result was equivalent to the inequality that Chebyshev derived for "general" random variables in 1867, which has been used ever since for deducing weak laws of large numbers.

With the accounts just mentioned, Laplacian error theory gained a purely mathematical quality beyond the practical necessities of error calculus. In this sense, asymptotic error theory became one of the roots of modern probability theory.

5.7 Chebyshev, Markov, and the Theory of Moments

Another strand of development toward modern probability was built up from Chebyshev's contributions, as well as from Markov's pre-1900 work. Only three papers by Chebyshev on probability reached a broader audience: The first two, published in 1846 and 1867, were devoted to Poisson's weak laws of large numbers. Chebyshev criticized Poisson's way of reasoning, which resorted to arguments concerning asymptotic normal laws for the distributions under question without precisely determining the errors of the applied approximations. In the first paper he gave rather sharp bounds for the probabilities of relative frequencies in a Bernoulli process with varying success probabilities. In the second he proved—using essentially the same arguments as Bienaymé—the now so-called "Bienaymé–Chebyshev inequality" in the form:

$$P\left(\left|\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mathrm{E}X_k\right| \leq \alpha \sqrt{\sum_{k=1}^{n} \mathrm{E}X_k^2 - \sum_{k=1}^{n} (\mathrm{E}X_k)^2}\right) > 1 - \frac{1}{\alpha^2}$$

for independent random variables X_k (Chebyshev explicitly considered the case of variables with a respectively finite number of values only) and positive α.

We do not know whether in 1867 Chebyshev actually knew Bienaymé's achievement. Later on, however, he even maintained that his reading of Bienaymé's pertinent article (which also contained a discussion of general even order moments of errors as measures of precision) had raised his interest in moment problems. In this respect, Chebyshev's chief problem was about estimates of linear mass distributions $\int_a^x f(t)\,dt$ from the knowledge of finitely many moments $m_k = \int_a^b f(t)\,t^k dt\,(k = 0, 1, 2, \ldots, n)$, where f denotes a non-negative mass density (intuitively conceived as delta-like in the case of discrete mass distributions) on an interval $[a, b]$. In 1874 Chebyshev published (without proof) inequalities from which such estimates resulted. During the next two decades, Chebyshev and his disciples—most notably Markov, who also published a proof of Chebyshev's original inequalities in 1884—refined and extended this newly established analytic theory of moments. (Outside Russia, Stieltjes especially, who had started pertinent work in the 1880s without knowledge of Chebyshev's activities, contributed to moment theory.)

Chebyshev's approach to the CLT should be viewed mainly in light of his moment activities. He stated the integral version of the CLT in its modern form, as a limit theorem and no longer as an assertion about an approximate normal distribution. $X_1, X_2, \ldots$ designating (tacitly independent) random variables (*quantités*) with zero expectations, Chebyshev presupposed that $\left|\mathrm{E}X_k^m\right| < C_m$, where C_m were given constants only depending on m. Markov later added the assumption that $\mathrm{E}X_k^2$ was always above a fixed positive bound

from a certain k on. Under these conditions, the CLT was stated in the form

$$P\left(\alpha\sqrt{2\sum_{k=1}^{n}\mathrm{E}X_k^2} \le \sum_{k=1}^{n}X_k \le \beta\sqrt{2\sum_{k=1}^{n}\mathrm{E}X_k^2}\right) \to \frac{1}{\sqrt{\pi}}\int_{\alpha}^{\beta} e^{-x^2}\,dx\ (n\to\infty). \qquad (5.3)$$

In his proof, Chebyshev used what we would now call the Laplace transform of the sum of random variables rather than its Fourier transform, and in this way he involved moments. He made it plausible that the limit distribution of the normed sum $\sum X_k/\sqrt{2\sum \mathrm{E}X_k^2}$ had the same moments as the normal distribution on the right side of (5.3), and, by aid of a theorem he had proven previously, he concluded the CLT. In principle, Chebyshev did not significantly go beyond Poisson's "approximation arguments," and he did not reach the aim he had established for himself of finding usable estimates for the deviation of the two sides in (5.3). In 1898–99 Markov revisited Chebyshev's proof and cautiously criticized its deficits. He rigorously showed that 1) the moments of each order of the normed sum tend to those of the normal distribution, and 2) convergence of moments implies convergence of distribution. Through Chebyshev's and Markov's work the CLT became an intra-mathematical issue independent of applications beyond mathematics. On the other hand, however, it still played a subservient role: for illustrating the power of the analytic theory of moments.

5.8 Kinetic Theory of Gases

Starting with a contribution by August Carl Krönig, the later so-called "statistical conception" of the kinetic theory of gases developed during the second half of the 19th century. In the simplest case a gas was conceived as a system of a "great many" little balls of negligible radius which perform elastic collisions with each other and with the vessel's wall.

The characteristics of the "statistical conception" were (i) more or less specific assumptions, intuitively made by referring to "rules of probability calculus," e.g., on independence of motions of the individual particles, or, on spatial homogeneity or isotropy within the systems under consideration, and (ii) a scarcely existing conceptual differentiation between relative frequencies and probabilities (corresponding to the "statistical approach" of the kinetic theory of gases). From the mathematical point of view, in almost all cases elementary combinatorial and asymptotic methods were sufficient for mastering the "technical" problems of the kinetic theory of gases. Still, the development of later so-called "statistical physics" corresponded to the general universalism toward applications as it was propagated by Laplace's "Essai philosophique."

Through contributions by Rudolph Clausius, who also introduced the notion of "mean free path," and through numerous works by James Clerk Maxwell and Ludwig Boltzmann—to refer only to the three most important proponents of the kinetic theory of gases—this discipline evolved in the direction of ever more refined and at the same time ever more generally applicable models, far beyond the simple "ideal gases."

Among Maxwell's most notable achievements were the velocity distribution named after him, and his concept of what is now called the "ensemble average." In his first derivation

of the velocity distribution, published in 1860, Maxwell concluded by arguments based on symmetry and stochastic independence of spatial directions that among a "great" total number N of particles the "number" dN of those with orthogonal velocity components between x and $x+dx$, y and $y+dy$, z and $z+dz$ was given by

$$dN = Nf(x)\,f\left(y\right)f(z)\,dx\,dy\,dz, \tag{5.4}$$

where, e.g.,

$$f(x) = \frac{1}{\alpha\sqrt{\pi}}\,e^{-x^2/\alpha^2} \tag{5.5}$$

with a constant $\alpha > 0$.

At the beginning of his account (and only there), Maxwell referred to the quantity dN in (5.4) as an "average number." However, in so doing he did not clarify the meaning of this term. Maxwell gave a precise analysis of those averages only in his last published article from 1879, basing his account on ideas which were in principle due to Boltzmann. Maxwell considered a "great" number N of gas systems, each containing the same "great" number of particles and each having the same total amount of energy E. The temporal evolution of every system was characterized by a $2n$-dimensional function of time t (n a very large number), the so called "path"

$$\left(q_1(t),\ldots,q_n(t);p_1(t),\ldots,p_n(t)\right) =: \left(q(t),p(t)\right),$$

$q_r(t)$, $p_r(t)$ indicating single coordinates of position and momentum. For any fixed time t_1, the $2n$-tuple $\left(p(t_1),q(t_1)\right)$ was termed a "phase." The method Maxwell called "statistical investigations" was based on the idea of calculating the later so-called (the designation is due to Gibbs) "ensemble average"—depending in general on the time t_1—of the systems having a phase within an infinitely small neighborhood of the phase $\left(p(t_1),q(t_1)\right)$, by establishing an appropriate relation between the "number" $dN(t_1)$ of these particular systems and the number N of "all" systems of the largest imaginable variety of initial values. From the point of view of the "late" Maxwell, "average" referred not to a number of individuals having a common state within one single system, but to the number of those gas systems having a certain common phase among "all" possible systems.

For a closer determination of relations between dN and N like (5.4) and (5.5), Maxwell assumed the system to have the property that all points of phase space (the space of all $2n$-tuples) which are compatible with the given total energy E are incident, at a particular moment of time, with the path of the system. In other words, this means that a system can reach any (in principle) possible state at a particular moment of time. Boltzmann later called this property "ergodicity." In most cases Maxwell and Boltzmann considered ergodicity to be more or less self-evident.

To Boltzmann we owe a far-reaching generalization of Maxwell's distribution for kinetic energies to systems with arbitrary underlying potential energies. An especially important feature of Boltzmann's work is that he introduced purely probabilistic considerations into the kinetic theory of gases, by conceiving gases as abstract urn models. In this way he established a methodological framework which could also be used in later quantum statistics. Boltzmann's combinatorial approach becomes especially apparent in his discussion of the "H-function," as summarized here following Ehrenfest's (1911/1959) exposition: Let a_i be the number of those gas particles within a certain system whose generalized coordinates of

position and momentum $(q_1, \ldots, q_s; p_1, \ldots, p_s)$ correspond to a point within a very small cell ω_i of the $2s$-dimensional space (s denoting the number of degrees of freedom). Then the H-function of the system's state Z characterized by the a_i is defined by

$$H(Z) = \sum a_i \log a_i.$$

If one calculates the probability for any Z under the constraint of a fixed total energy by considering the different combinations in distributing molecules among cells, and if the volumes of the sets ω_i are assumed to be vanishing and the number of particles is assumed to be "very large," then the Maxwell-Boltzmann distribution corresponds to a state Z_0 which is the most probable. On the other hand, $H(Z)$ becomes a minimum among all possible states Z if $Z = Z_0$. Boltzmann made it plausible that, during the temporal evolution of a gas system, the corresponding H-function would be very close to its minimum with "overwhelming probability."

At many places Boltzmann used the word "probability" in a rather intuitive way without precisely clarifying this notion. Relating to the probability that the phases (p, q) of a gas system lie within a set ω of the energy surface (the manifold of all phases compatible with the system's total energy), Boltzmann preferred the concept of "time average:" the sum of all periods during which the system has its phase within ω divided by the total observation period T, where $T \to \infty$ is assumed. Ergodicity implied equality of ensemble average and time average.

Around the end of the 19th century, in the field of kinetic theory of gases a puzzling variety of notions of probability existed: probabilities according to time average and to ensemble average, relative frequencies interpreted as probabilities, for example with respect to the "number" of particles with velocities between v and $v + dv$, "combinatorial" probabilities in relation to the H-function. The mathematical foundations of and the relations between these notions were widely considered to be unclear, not to mention the open problem of the ergodic hypothesis. Therefore, it is not surprising that Hilbert, when including axiomatics of probability theory as the 6th problem in his famous list of the 23 most important mathematical problems in 1900, explicitly referred to problems concerning "average values" in the kinetic theory of gases.

The kinetic theory of gases also had a decisive impact on the philosophical discussion of the notions of randomness and probability. Laplacian determinism was gradually replaced by a kind of semi-determinism emphasizing the fact that, despite complete knowledge of all underlying physical laws, even an immeasurable variation of initial conditions could induce a totally different process, which therefore appeared to be entirely random. The probabilities of those processes would have an objective character, however, because of their accessibility to calculation in a physical model or because of the possibility of determining the probabilities from observed relative frequencies. In this way, statistical physics essentially influenced the transition from a deterministic to an indeterministic world view, which emerged surrounding the development of 20th-century quantum theory.

Initially, further progress in statistical physics only partially followed Hilbert's program, but it maintained its significance as a stimulus for the development of mathematical probability theory. In his 1902 book *Elementary principles in Statistical Mechanics*, Josiah Willard Gibbs gave a quite abstract framework for the statistical treatment of mechanical

systems, which was not sufficiently general, however, to be used for solving all essential problems completely satisfactorily. In 1911 Paul and Tatjana Ehrenfest provided a thorough analysis of the results previously achieved and the remaining problems of statistical physics, which was apparently strongly influenced by Hilbert's program and which reached the status of a classic. The mathematical discussion of Brownian motion, which began around 1905, had a crucial influence upon the theory of probability in function spaces as it was developed by Norbert Wiener from the 1920s on. As early as 1913, Rosenthal and Plancherel independently showed the impossibility of non-trivial ergodic systems. Further achievements addressing the problems concerned with ergodicity eventually led to a self-contained "ergodic theory" in the 1930s. Kolmogorov, in his "ultimate" 1933 approach to probabilistic axiomatics still explicitly referred to the influence of problems of statistical physics on his work.

5.9 The Emergence of Modern Probability

What is "modern probability"? When did it first arise? In order to answer this question we have to refer to criteria for "modern," and, in particular, for "modern mathematics" (e.g., Mehrtens 1990). An essential characteristic of modern mathematics is its chief goal of exploring all possibilities independent of external aspects such as "applicability," "utility," "beauty," or "relations to real world." The "truth" and "value" of mathematical work are dependent only on self-imposed rules. In this respect, modern mathematics is a self-referential and independent system.

Around the turn of the century, probability theory was conceived as a part of the natural and moral sciences rather than of mathematics proper. There had been work on probability in an abstract scope already, e.g., by Chebyshev, but, as we have seen, full independence was not yet achieved with these contributions. One instance of full independence is Lyapunov's work on the CLT in 1900/01. Lyapunov, despite being a representative of the "St Petersburg school," which had been founded by Chebyshev, dispensed with a focus on moments in order to concentrate upon the greatest generality of the assumptions for the CLT, with the goal of sounding out the possibilities of this theorem as far as possible.

As pointed out by von Plato (1994), another instance is Borel's work on denumerable probabilities in 1909, where a problem alien to classical probability was discussed: What is the probability of infinitely many successes in a Bernoulli process with infinitely many trials? This problem could not be conceived as a "natural" extension of a "real" situation (e.g., a game with independent repetitions), but was based on "modern" concepts of infinite sets. During the first decades of the 20th century, the point of view of probability as an abstract measure on abstract sets developed step by step and thus strongly contributed to the growing separation of mathematical theory and applications.

A third aspect of modern probability was the goal of weakening the supposition of independence in studying repeated trials. Markov started with such investigations in a 1907 paper on a generalization of the weak law of large numbers. Apparently, his chief motivation was to extend "classical" results as far as possible from the intra-mathematical point of view.

5.10 Axiomatics and Inclusion of Measure Theory

As described above, Hilbert in 1900 demanded that probability theory be established in an axiomatic way. He referred to an attempt in the same year by Georg Bohlmann, whose "axioms," however, were hardly more than a list of the usual basic rules of probability. Until the beginning of World War I a few authors contributed to Hilbert's problem on the basis of contemporary set and measure theory; but they still adhered very strongly to the paradigm of equiprobability. In 1905, Émile Borel, hinting at a probabilistic problem concerning continued fractions, due to Torsten Brodén and Anders Wiman (who had published work on this topic around 1900), briefly described—in the framework of real subsets—the idea of restricting oneself to measures of later so-called "Borel sets". These are real sets generated by the union or intersection of countably many intervals, and Borel had introduced them in 1895 in his thesis on the theory of functions.

Richard von Mises (1919b) tried to base probability theory on two axioms only. He considered so called *Kollektivs*, i.e., sequences of "objects of imagination" to each of which a k-tuple of real numbers was assigned as a *Merkmal* ("label"), where at least two different labels should occur infinitely many times within a *Kollektiv*. He postulated the existence of *Kollektivs* such that: 1) for any $A \subset \mathbb{R}^k$ the relative frequency of the event that a label in A occurs in the sequence under consideration tends to a limit, and 2) for each subsequence which is chosen "without considering any distinctions regarding the labels" the same limit of the relative frequency assigned to A exists. Von Mises apparently regarded *Kollektivs* as ideal schemes, whose existence was self-evident. In 1937 Abraham Wald showed the formal consistence of von Mises's (somewhat modified) axioms, but it was only in the 1960s that models with an entirely satisfying "random" behavior were established.

Von Mises (1919a) also introduced the standard of treating real and vector-valued random variables (or the respective *Kollektivs*) by their distribution functions. He defined a distribution function F on $\mathbb{R}$ as a monotonically increasing function, continuous on the right, with $F(-\infty) = 0$ and $F(\infty) = 1$. $F(x)$ can be interpreted as the probability $P(X \leq x)$ of a real valued random variable X. Using the device of Stieltjes integral (which had originally been developed in 1892 in the context of moment problems, and was well established by 1920), for $r > 0$ the expected value of X^r (if it exists) can be expressed by $\int_{-\infty}^{\infty} x^r dF(x)$. Convergence of a sequence F_n of distributions to a distribution F can be defined by the condition that, for each point x in which $F(x)$ is continuous, $F_n(x) \to F(x)$. The distribution function F_{12} of the sum $X_1 + X_2$ of two real valued independent random variables with distribution functions F_1 and F_2 is given by the convolution

$$F_{12}(x) = \int_{-\infty}^{\infty} F_1(x-y)\, dF_2(y) =: F_1 \star F_2(x).$$

Similar properties are valid in the multi-dimensional case. Von Mises already used the now common word *Faltung*, even though not for the distribution function of the sum but for the corresponding combination of two *Kollektivs*. Through his distribution functions, von Mises established a connection between probability, measure, and integration. With

his frequentist approach, however, he retained a strong relation to the "world of experience" at the same time.

Already in 1913, Johann Radon had expounded a theory of Stieltjes integrals in $\mathbb{R}^n$ in the sense of an extension of Lebesgue integrals. In 1915, Maurice Fréchet generalized Lebesgue integrals to integrals with respect to a sigma additive (i.e., countably additive) measure on abstract sets constituting a sigma algebra (i.e., an algebra consisting of certain subsets of a given set Ω—where it is required that Ω is itself an element of the algebra—such that, for any two of these subsets, their set-difference is an element of the algebra, and, for any countable family of such subsets, its union is an element of the algebra as well). In 1924 Fréchet's lectures on this topic appeared, and in this work many open questions of the original contribution were settled, such that this account would have been sufficient, in principle, to give the measure theoretic foundation for establishing probability theory on an abstract sigma algebra. (See Hochkirchen 1999, section 2.2, and Shafer and Vovk 2006 for details.)

It was not until 1933, however, that Andrei Nikolaevich Kolmogorov did publish axioms which have remained the standard ever since. Probability is a measure on a sigma algebra with values in [0,1] and with properties equivalent to sigma-additivity. Because of its totally abstract orientation toward sets and measures, Kolmogorov's account could not establish any relations to the "real world." With regard to this issue, he referred to von Mises's approach. Indeed, Kolmogorov's system of axioms essentially developed from the mathematical folklore of around 1930, and with this achievement alone, his 1933 work would not have reached its paramount significance. From today's point of view the true importance of Kolomogorov's contribution was due to his definition of random variable as a measurable mapping from one sample space into another, conditional probability and expectation on the basis of the theorem of Radon-Nikodym, and his "fundamental theorem," the now so-called "consistency theorem" guaranteeing the unique existence of probability distributions in particular probability spaces of infinite dimension if the respective finite-dimensional marginal distributions are given.

At the time, the reception of Kolmogorov's account was rather restrained. Contemporary work on real or vector valued random variables did not require the full generality of his exposition. Investigations of stochastic processes in the sense of random elements in function spaces were, with the exception of Wiener's work on Brownian motion, in a very early stage of development, and the application of Kolmogorov's consistency theorem to these problems was not possible without significant additional effort. The start of the true success story of Kolmogorov's *Grundbegriffe* was around 1950. Only from this time were its concepts broadly appreciated in their full generality, and further developed.

5.11 Convergence of Distributions: The Further Development of "Classic" Limit Theorems

The CLT and the weak law of large numbers are the only significant subjects of modern probability theory with roots in classical probability. As already mentioned, Lyapunov's work on the CLT in 1900–01 can be considered an important step into modern probability.

Basing his work essentially on Poisson's methods, Lyapunov showed that the CLT for independent random variables with zero means in the sense of (5.3) holds if

$$\frac{(d_1 + d_2 + \cdots + d_n)^2}{(a_1 + a_2 + \cdots + a_n)^{2+\delta}} \to 0 \quad (n \to \infty),$$

where $d_i := \mathrm{E}|X_i|^{2+\delta}$ ($\delta > 0$ arbitrarily small), and $a_i := \mathrm{Var}X_i$. In proving the CLT, Lyapunov also derived a uniform upper bound for the absolute value of the difference between the exact probability under question and the approximate probability according to the normal distribution. Lyapunov's estimate was considerably refined in later works, by Harald Cramér in 1923, and especially by Andrew C. Berry in 1941, and Carl Gustav Esseen in 1942–3. (See Fischer 2011, section 5.2.8 for details.)

In 1908, Markov succeeded in reducing Lyapunov's CLT to "his" and Chebyshev's CLT (which had been deduced by moment methods) by introducing a new device which proved to be especially fruitful in the context of limit theorems during the next decades: the truncation of random variables. Given a monotonic sequence of numbers N_n tending to infinity, he considered, for $k \leq n$, instead of the random variables X_k (without loss of generality assumed to have zero means), the truncated variables

$$X'_{nk} := \begin{cases} X_k \text{ for } |X_k| \leq N_n \\ 0 \quad \text{for } |X_k| > N_n . \end{cases}$$

He showed that, under Lyapunov's condition, a sequence N_n can be found such that

$$\sum_{k=1}^{n} P(X'_{nk} \neq X_k) \to 0 \ (n \to \infty), \tag{5.6}$$

and the truncated variables are accessible to a derivation of the CLT via moments. (5.6) implies that the CLT is likewise valid for the un-truncated variables. (See Fischer 2011, section 5.1.5.)

In 1919, von Mises did not only expound his "statistical" concept of probability, he also gave a comprehensive account on characteristic functions (named *komplexe Adjunkte* by him) which strongly influenced other authors during the 1920s. Notwithstanding slight deviations in the particular usage of individual authors, the characteristic function $\varphi(t)$ (the name is probably due to Poincaré) of a random variable X with distribution function F is defined by

$$\varphi(t) := \mathrm{E}e^{itX} = \int_{-\infty}^{\infty} e^{itx} dF(x) \tag{5.7}$$

for $t \in \mathbb{R}$. The most important property of characteristic functions—besides their unrestricted existence—is that, for independent variables X_1 and X_2, the characteristic function of $X_1 + X_2$ is the product of the characteristic functions of X_1 and X_2, respectively. Therefore, characteristic functions are especially appropriate for investigating stochastic properties of sums of independent random variables. Von Mises showed that, for random variables with densities as well as for variables taking only equidistant discrete values, convergence of characteristic functions implies convergence of distributions, and he also derived quite general local versions of the CLT in a similar way. Concerning the CLT for variables with general distribution functions, however, he was not able to apply the method of characteristic functions successfully. This latter problem was solved by Paul Lévy in 1922.

He proved that, under rather general conditions, convergence of distributions $F_n \to F$ corresponds to the convergence of the respective characteristic functions $\varphi_n(t) \to \varphi(t)$ for all real t. In 1937, Cramér was able to further weaken Lévy's assumptions for the "continuous correspondence" of characteristic functions and distributions.

Likewise in 1922, Jarl Waldemar Lindeberg published his epochal article on the classical CLT asserting

$$P\left(\sum_{k=1}^{n}(X_k - \mathrm{E}X_k)\Big/\sqrt{\sum_{k=1}^{n}\mathrm{Var}\,X_k} \le x\right) \to \Phi(x),$$

where Φ designates the distribution function of the standard normal distribution, under the now so called "Lindeberg condition" which later proved to be necessary if the individual random variables were asymptotically negligible in a certain sense. For his proof, Lindeberg did not apply characteristic functions. Instead, he derived an entirely new method of proof, which, in principle, consisted of elementary considerations on the convolution of the distribution functions of the single random variables and an additional auxiliary variable whose influence on the sum could be assumed to be arbitrarily negligible. Lindeberg himself had originally expounded his condition for the CLT in two equivalent versions, which are widely forgotten now. The later commonly used (and likewise equivalent) "Lindeberg condition," however, cannot be found in his work. It was stated by Lévy, and has (assuming zero expectations) the form:

$$\frac{1}{r_n^2}\sum_{i=1}^{n}\int_{|x|>tr_n} x^2\,dF_i(x) \to 0 \quad \forall t > 0,$$

where F_i designates the distribution of the i-th summand, and r_n^2 the variance of the sum.

Around 1925, innovative versions of the CLT appeared in the work of Lévy and Bernshtein. The respective generalizations referred to a new mode of norming, by considering sums $\sum_{k=1}^{n}(X_k - a_k)/b_n$ of independent random variables, where the constants a_k and b_n were, in general, different from expectations and standard deviations whose existence was no longer required. Furthermore, the assertion of the CLT was extended to stable limit distributions, i.e., to distributions V with the property that for all positive a, b there exists a constant c such that $V(a\cdot) \star V(b\cdot) = V(c\cdot)$ ($V(a\cdot)$ designates the distribution function $x \mapsto V(ax)$). Characteristic functions proved to be especially powerful in discussing stable distributions.

Around 1935, the "traditional" CLT (with general norming, however) reached its "ultimate" form. Lévy and William Feller almost simultaneously but independently published sufficient and even necessary conditions in order that the distributions $P\left(\sum_{k=1}^{n}(X_k - a_k)/b_n \le x\right)$ of normed sums of independent random variables tend to the standard normal distribution, under the general assumption of uniform smallness of all random variables under consideration, expressed, e.g., by the condition $\max_{1\le k\le n} P(|X_k - a_k| > \varepsilon b_n) \to 0$ for all $\varepsilon > 0$. Lévy's and Feller's accounts were quite different concerning methods and style and provided different but equivalent conditions for the CLT. Whereas Feller used characteristic functions exclusively, Lévy employed his newly derived devices of concentration and dispersion. For positive l the concentration $f_X(l)$ of a random variable X is given by $f_X(l) = \sup_{a\in\mathbb{R}} P(a < X < a + l)$, whereas for any $\gamma \in (0,1)$ the dispersion $\varphi_X(\gamma)$ is given by $\varphi_X(\gamma) = \inf\{l \in \mathbb{R}^+ | f_X(l) \ge \gamma\}$. Expressed in Lévy's language of dispersions,

the above-mentioned necessary and sufficient condition for the CLT (under the general assumption of zero medians for all random variables) is:

$$\exists \gamma \in (0,1)\ \forall \varepsilon > 0 :\ P\Big(\max_{1 \le k \le n} |X_k| > \varepsilon \varphi_{\sum_{k=1}^{n} X_k}(\gamma) \Big) \to 0 \ \ (n \to \infty).$$

Eventually, until 1940, the assertion of the CLT was further generalized (especially in Gnedenko's and Doeblin's pertinent work), by considering infinitely divisible distributions as limit distributions for normed sums. Infinitely divisible distributions had emerged from investigations concerning stochastic processes with independent increments, starting with contributions by de Finetti in 1929. Originally, infinitely divisible distributions V were defined by the property that, for any natural n, specific distributions $V_{n1}, \ldots, V_{nn}$ existed such that $V = V_{n1} \star \cdots \star V_{nn}$. In 1937, it became clear through pertinent work by Khinchin that this property was equivalent to the (now commonly used) characteristic that, for each natural n, a distribution F_n exists such that $V = F_n \star \cdots \star F_n$ (n-times). In 1934 Lévy, and in 1937 Khinchin found formulae for characterizing infinitely divisible distributions by their characteristic functions. (See Fischer 2011, section 7.2 for details.)

Progress regarding the CLT was accompanied by progress with respect to the weak law of large numbers (5.2). The common derivation via the Bienaymé-Chebyshev inequality requires the existence of the variance for each of the mutually independent random variables, although second-order moments do not occur in the formulation of this law. In 1913, Markov had already abandoned the assumption of finite variances, and using truncated variables he proved the validity of (5.2) under the condition that for all variables $\mathrm{E}|X_i|^{1+\delta}$ remains below a uniform upper bound for an arbitrarily small $\delta > 0$. The assertion of the weak law of large numbers was eventually extended to general norming, in the form $P\left(\left|\sum_{k=1}^{n} (X_k - a_k)/b_n\right| > \varepsilon\right) \to 0$. Kolmogorov in 1928, and in a significantly more explicit way Feller in 1937, achieved general results in this respect in the sense of necessary and sufficient conditions.

By 1940, the theory of limit distributions for sums of independent random variables had at least roughly reached its final form. The book by Gnedenko and Kolmogorov on this topic, whose first Russian edition appeared in 1949 and which essentially summarized the results obtained up to the beginning of World War II, became one of the most important and influential monographs of 20th-century probability theory. The field of problems surrounding distributions of sums of independent random variables is an impressive historical example of how traditional contents with essentially traditional methods were transformed into an extremely modern part of mathematics, a process which was basically motivated by the aim of sounding out all mathematical possibilities even beyond any criteria of convenience and applicability.

5.12 Strong Laws

In his above-mentioned 1909 paper, Borel considered an infinite number of independent trials, each with at most denumerably many outcomes. In the case of an infinite process with only two outcomes in each trial, "success" and "failure," and (in general different) success

probabilities $p_1, p_2, \ldots$, Borel (by partially questionable arguments) showed that, for any natural m, the probability of more than m successes is one if $\sum_{j=1}^{\infty} p_j = \infty$, whereas the probability for infinitely many successes is zero if $\sum_{j=1}^{\infty} p_j < \infty$. In the same paper Borel also provided a first version of the strong law of large numbers for the relative frequency in a Bernoulli process in the context of the question about the "probability" that the proportion of zeros and ones in the dyadic representation of a real number is $1/2$. Borel, in an admittedly "partially arbitrary" way, assumed the dyadic expansion of any real number generated by an infinite Bernoulli process with success probability $1/2$ for zeros and ones. He simply maintained that his results could also be proven by considerations regarding Lebesgue measures, which principle he had already hinted at in his above-mentioned 1905 paper. In this way Borel stimulated subsequent work by different authors on measure theoretic aspects of number theory, which, in turn, inspired general measure theoretic concepts in probability theory.

In 1917, Francesco Cantelli proved a fairly general strong law of large numbers for the arithmetic mean $S_n/n := \left(\sum_{k=1}^{n} X_k - \mathrm{E}X_k\right)/n$ of independent random variables X_k, where the sum of the fourth order central moments $\sum_{k=1}^{n} \mathrm{E}(X_k - \mathrm{E}X_k)^4$ was assumed not to grow faster than n. With the aid of a generalization of the Bienaymé-Chebyshev inequality toward moments of fourth order, he essentially showed that, with an appropriately chosen sequence $a_n > 0$ tending to zero,

$$P\left(\sup_{i\in\mathbb{N}} \left|\frac{S_{n+i}}{n+i}\right| > a_n\right) \to 0 \quad (n \to \infty). \tag{5.8}$$

The most spectacular results on strong laws between ca. 1925 and 1930 were achieved by Kolmogorov and Aleksandr Yakovlevich Khinchin (to whom the designation "strong law" seems to be due). These results included strong laws for arithmetic means of independent random variables in the sense of (5.8) (e.g., if $\sum_{k=0}^{\infty} \mathrm{Var}X_k/k^2 < \infty$, Kolmogorov in 1930), theorems on the probability 0 or 1 for the convergence of series of independent random variables (Kolmogorov and Khinchin in 1925, Kolmogorov in 1928), and estimates of the velocity of convergence in strong laws. The last problem resulted in the "law of the iterated logarithm," which was derived by Khinchin in a 1924 paper for Bernoulli processes, and later significantly generalized by Kolmogorov in 1929: Let X_k be a sequence of uniformly bounded independent random variables with zero expectations and non-degenerate distributions, $S_n := \sum_{k=1}^{n} X_k$, $B_n := \mathrm{Var}S_n$. In order that the two relations

$$P\left(\exists k \geq n_0 : |S_k| > (1+\delta)\sqrt{2B_k \log(\log(B_k))}\right) < \varepsilon,$$
$$P\left(\exists k \geq n_0 : |S_k| < (1-\delta)\sqrt{2B_k \log(\log(B_k))}\right) > 1-\varepsilon \tag{5.9}$$

hold for arbitrarily small positive ε and δ and for arbitrarily large natural n_0, it is necessary and sufficient that $B_n \to \infty$. In 1928 Khinchin also generalized strong laws for arithmetic means to non-independent random variables, in assuming their mutual correlation coefficients to be bounded in a certain sense (for details on strong laws see Hochkirchen 1999, section 8.1, and Khinchin 1933).

At this place only a few hints at basic ideas underlying strong laws are possible: The concept of so called "equivalent" sequences of random variables X_k and Y_k, defined,

according to Khinchin, by the property $\sum_{k=1}^{\infty} P(X_k \neq Y_k) < \infty$ became a frequently used device. Kolmogorov's inequality

$$P\left(\max_{1\leq k\leq n}\left|\sum_{i=1}^{k} X_i\right| \geq R\right) \leq \sum_{k=1}^{n} \mathrm{E}X_k^2 \big/ R^2 \quad (R > 0)$$

for a sequence of independent random variables X_k with zero expectations, was proven in 1928 and has served as a basis for many proofs since that time. The significance of sigma additivity was especially clearly visible in the context of strong laws.

Fréchet, apparently motivated by the striking analogies between probability and random variables (which he still treated in a rather intuitive way) on the one hand, and Lebesgue measure and measurable functions on the other, delivered in 1930 a comprehensive study of several modes of convergence and their mutual relations: "Convergence in probability" (stochastic convergence, as in weak laws of large numbers), "almost sure" or "strong" convergence, and convergence of distribution functions. An ultimate clarification of "strong convergence" was provided by Kolmogorov on the basis of his notion of random variable in the *Grundbegriffe*: If Ω and $\mathcal{A}$ designate the sample space and the sigma algebra on which the probability measure P is defined, and if $X_n : \Omega \to \mathbb{R}$ is a sequence of random variables, then the set $\{\omega \in \Omega | X_n(\omega) \text{ converges}\}$ is an element of $\mathcal{A}$. On this conceptual basis, the "finitary" assertions (5.8) and (5.9) can be shown to be equivalent to $P(\lim_{n\to\infty} S_n/n = 0) = 1$ and $P\left(\limsup_{n\to\infty} |S_n| / \sqrt{2B_n \log(\log(B_n))} = 1\right) = 1$, respectively.

5.13 Weakening Independence

Apparently, the main goal of Markov's first investigation, in 1907, of a random experiment with non-independent trials was his interest in generalizing conditions for the weak law of large numbers. He considered repeated trials with only two outcomes A and A' such that, in each trial, $P(A) = p$, and the occurrence of A or A' depended on the outcome of the immediately preceding trial only. Markov was able to derive inequalities of the Bienaymé-Chebyshev type, from which the weak law of large numbers for relative frequencies could be deduced. In a supplement to the same paper, Markov extended the result to a chain of random variables X_k, each with the same set of finitely many values such that X_k was dependent only on X_{k-1} with transition probabilities independent of k, and he proved that $\sum_{k=1}^{n} X_k/n$ stochastically converges to $\mathrm{E}X_1$. In the following years, Markov generalized those results, especially by extending the assertion of CLT via moment methods to sums of chained variables under a variety of specific conditions.

An important step to further generalizations of dependence was made by Sergei Natanevich Bernshtein, by the publication of a brief note on a CLT for normed sums of "almost independent" random variables in 1922. (At the beginning of the paper Bernshtein noticed that he had obtained the results in 1917–18 already.) A comprehensive article with proofs and further applications was delivered in 1926. If S_n denotes the sum of the first n variables X_k (all with zero means), and B_n the second order moment of the sum, Bernshtein presupposed that the absolute values of the conditional expectations of X_k, X_k^2,

and $|X_k^3|$, as dependent on $X_1, \ldots, X_{k-1}$, were always bounded from above by α_k, β_k, and c_k, respectively, and he assumed that $\sum_{k=1}^n \alpha_k \Big/ \sqrt{B_n}$, $\sum_{k=1}^n \beta_k \Big/ B_n$, and $\sum_{k=1}^n c_k \Big/ \sqrt{B_n^3}$ tend to 0 with $n \to \infty$. Bernshtein applied conditional expectations in a still "naive" way, apparently assuming only discrete random variables, each with only a finite number of values. His CLT, asserting $P(S_n \Big/ B_n \le x) \to \Phi(x)$ and named *lemme fondamental*, could also be applied, on the basis of a clever device, for investigating normed sums of random variables forming Markov chains in quite general settings.

The next decisive innovation was due to Lévy with his studies on martingale limit theorems from 1934 on. (See Fischer 2011, section 7.1.3 for details.) He considered a sequence of random variables (X_n) with the property

$$\mathrm{E}X_1 = 0, \quad \mathrm{E}(X_n \,|\, X_1, \ldots, X_{n-1}) = 0 \quad (n = 2, 3, \ldots), \text{ and}$$

$$P\left(\sum_{k=1}^{\infty} \sigma_k^2 = \infty\right) = 1, \text{ where} \quad \sigma_k^2 = \mathrm{E}(X_k^2 | X_1, \ldots, X_{k-1}).$$

In contrast to Bernshtein, Lévy did not discuss "classical" sums but sums of the form

$$S(t) := \sum_{k=1}^{N(t)} X_k,$$

where $N(t)$ was itself a random variable, defined by

$$N(t) = \min\{n \in \mathbb{N} | \sigma_1^2 + \cdots + \sigma_n^2 \ge t\}.$$

In modern terminology, $S(t)$ can be interpreted as a martingale (this name was introduced by Jean Ville in 1939, as it appears). Lévy discussed various conditions for the assertion of the martingale CLT

$$\lim_{n \to \infty} P(S(t) < x\sqrt{t}) = \Phi(x)$$

which were, at least in part, substantially weaker than Bernshtein's.

5.14 Stochastic Processes

Probability distributions depending on a continuous time scale had already been introduced in the kinetic theory of gases, especially in relation to Boltzmann's equation from 1872 for the temporal evolution of the probability density of a particle with respect to its (generalized) coordinates of position and velocity. Nevertheless, only after the turn of the century did work on time-dependent distributions emerge in considerably increasing quantity and variety regarding problems and methods. Whereas, in the theory of gases, properties of the collective were in the foreground, now the main focus was on individual stochastic fluctuations.

In an elementary scope, a stochastic process with continuous time parameter is considered as a family of (one- or multi-dimensional) random variables $(X_t)_{t \in T}$ with distribution functions F_t, where $T = [t_0, b)$ is a bounded or unbounded interval.

In many cases, the transition from one moment of time t_1 and one single point $X_{t_1} = x (x \in \mathbb{R}^k)$ to an event $E_y := \{u \in \mathbb{R}^k \mid u_i \leq y_i,\ i = 1, \ldots, k\}$, $y = (y_1, \ldots, y_k)$, which happens at the time t_2, is given by a probability $P(t_1, x, t_2, E_y)$, such that the now so-called Chapman-Kolmogorov equation

$$P\left(t_1, x, t_2, E_y\right) = \int_{z \in \mathbb{R}^k} P\left(t_3, z, t_2, E_y\right) dP_z(t_1, x, t_3, E_z)$$
$$\forall t_1 < t_3 < t_2 \in T, x, y, z \in \mathbb{R}^k \tag{5.10}$$

holds. Then, the distribution function F_t can be obtained by

$$F_t(z) = \int_{\mathbb{R}^k} P(t_0, x, t, E_z)\, dF_{t_0}(x).$$

In particularly important applications the transition probabilities $P(t_1, x, t_2, E_y)$ depend not on the specific times t_1, t_2 but only on their difference Δt. These processes are called "diffusion processes" (or processes with time-homogeneous increments). Another special type according to (5.10) consists of stochastic processes with "independent increments," i.e., processes in which the increments $X_{t_2} - X_{t_1}$ are independent of X_τ for all $t_1 < t_2$ and $\tau \leq t_1$ $(\tau, t_1, t_2 \in T)$. If a diffusion process has independent increments, then the transition probabilities depend neither on the particular positions x, y nor on the particular moments of time t_1, t_2 but only on their differences $\Delta x = y - x$, and $\Delta t = t_2 - t_1$. Brownian motion is the most prominent instance of such a process.

In 1931 Kolmogorov provided a general theory of stochastic processes, in which problems related to equations (5.10)—even in a more general setting—played a major role. This work was one of Kolmogorov's decisive steps toward his measure theoretic exposition of probability theory in 1933. Since 1900, however, special diffusion processes, and, in particular, Brownian motion, had been studied. In his 1900 doctoral thesis, Louis Bachelier modeled the temporal behavior of stock exchange rates $x \in \mathbb{R}$ with an equation for the densities of transition probabilities $f(\Delta t, \Delta x)$ corresponding to:

$$f(t + \Delta t, z) = \int_{-\infty}^{\infty} f(\Delta t, z - x) f(t, x)\, dx. \tag{5.11}$$

By a guess, a solution was achieved in a form equivalent to

$$f(\Delta t, \Delta x) = \frac{1}{a\sqrt{2\pi\, \Delta t}} \exp\left(-\frac{(\Delta x)^2}{2a^2\, \Delta t}\right). \tag{5.12}$$

In the following section of his thesis, Bachelier provided a second method for obtaining a solution, the consideration of a (later so-called) random walk. He arranged equidistant points $0, \delta t, 2\delta t, \ldots$ on the time scale and assumed that in each of these moments stock exchange prices increased independently of all preceding changes by $\pm \delta x$ with the respective probability 0.5. Setting $\delta x = a\sqrt{\delta t}$ and assuming $\Delta t \big/ \delta t$ to be a very large natural number, he derived a relation equivalent to (5.12) by a plausibility consideration, based on de Moivre's theorem.

Shortly after Bachelier, Filip Lundberg in 1903 started work on stochastic processes with independent increments in the context of his discussion of the risk underlying insurance business. Thus, financial mathematics apparently was earlier than statistical physics, which

however, after Albert Einstein's mathematical treatment of Brownian motion in 1905, developed as the main field of activity in stochastic processes until ca. 1930. From a functional equation analogous to (5.11), Einstein derived (if in a rather intuitive and mathematically un-rigorous way) the homogeneous parabolic equation

$$\frac{\partial}{\partial t} f(t,x) = D \frac{\partial^2}{\partial x^2} f(t,x) \ (D > 0),$$

from which he obtained a solution corresponding to (5.12) (with $a^2 = 2D$) for the probability density of a linear Brownian motion departing from $x = 0$ at $t = 0$. Only one year later, Marian von Smoluchowski modeled a three-dimensional Brownian motion by a random walk in which, after having passed a linear segment whose length is the mean free path, a deflection with a certain fixed angle but a random spatial direction (uniform probability distribution) occurs at each step. In 1913, the same author introduced, in his investigation of Brownian motion under the impact of an external force, an analog to (5.10) for the evolution of the probability density in a diffusion process with increments dependent on the respective positions of the particles. (For details of von Smoluchowsky's work see Brush 1976 vol. 2.) Beside von Smoluchowski's contributions, a good deal of pertinent work by other authors appeared until 1920, for example Langevin, Fokker, or Planck, to mention but a few names. Around 1920, "stochastic fluctuations" had become established as a particular field of research in physics.

Processes governed by relations (5.10) have the "Markov property" that, for $t_0 < s < t$, the random variable X_t depends only on X_s and not on earlier states of this random process. Until ca. 1930, however, this property was brought only sporadically into a connection with Markov's original approach to dependent variables. And only around this time did the discussion of stochastic processes attain a purely mathematical quality. Now also Markov processes with discrete time parameter came to the fore, as they were studied in connection not only with random walks but also with mixing problems and problems concerning ergodic properties. In this way began a certain unification of Markov's original (predominantly intra-mathematical) approach, and extra-mathematical applications which had emerged without reference to Markov's ideas. Pertinent work included contributions by von Mises (who had already started studying ergodicity, around 1920) and his disciple Paul Höflich, Jacques Hadamard, Bohuslav Hostinský, and Georg Pólya. In Khinchin's summarizing exposition on "asymptotic laws" (1933) the method of deriving parabolic differential equations for distributions obeying (5.10) was applied to solving various problems, in particular on diffusion processes. It was also shown that, under the condition that the transition probabilities have sufficiently smooth densities (with respect to x and t), the distributions of diffusion processes with discrete time points converge to the distributions of the corresponding continuous time diffusion processes for each t, if the differences between the discrete time points tend to zero in a certain sense.

In connection with the investigation of the properties of infinitely divisible distributions a complete characterization of all distributions F_t of processes with independent increments could be achieved, at least in the one-dimensional case, mainly by the contributions of Kolmogorov, Lévy, Bavli, and Khinchin (see Fischer 2011, section 7.2.1). Lévy investigated the various properties of Brownian motion exhaustively until 1940, and he answered the most important questions concerning this particular issue.

Almost independently of the rest of the "mathematical world," Norbert Wiener pursued the "global" concept of random process as an element in an appropriate function space, beginning in the early 1920s. In the case of a linear Brownian motion within the time interval $[0,1]$ this function space is the "Wiener space" $\mathcal{C}$ of all continuous functions $x : [0,1] \to \mathbb{R}$ with $x(0) = 0$. Wiener succeeded in establishing a certain type of integral for specific functions $f : \mathcal{C} \to \mathbb{R}$ (e.g., continuous functions with respect to the L^∞ topology) over $\mathcal{C}$, where these integrals were consistent with the corresponding mean values assigned to the finite-dimensional marginal distributions of Brownian motion. Wiener's ideas would have been usable, in principle, for establishing a sigma additive measure on $\mathcal{C}$ according to Brownian motion. But his pertinent accounts remained largely unnoticed. Only a few American mathematicians (Cameron, Martin, Donsker, see Fischer 2011, section 7.3.1) would study properties of the Wiener integral in a more intensive way during the 1940s. On the other hand, despite first results by Doob, the application of Kolmogorov's consistency theorem to measures in function spaces remained in a rather provisional state until 1940. Only from ca. 1950 on did the study of stochastic processes develop within a complete and elaborated measure theoretic framework, as one can see, for example, from Doob's 1953 monograph.

An essential characteristic of the development of probability theory from 1900 to ca. 1940 is that its subdisciplines, which were at first essentially independent from one another regarding problems and methods, grew together, particularly during the decade 1930–40. A first completion of this process of integration becomes obvious from first monographs which appeared in the second half of the 1930s, e.g., by Fréchet (1936/38), Lévy (1937), or Cramér (1937), and went into further editions after 1950. World War II strongly impeded the further development of probability theory, which was, however, resumed immediately afterwards, and eventually led, until ca. 1960, to the shape of probability theory which is familiar today.

References

Bachelier, Louis (1900) *Théorie de la spéculation*. Paris: Gauthier-Villars.

Bernshtein, Sergei Natanovich (1922) Sur le théorème limite du calcul des probabilités. *Mathematische Annalen*. 85. pp. 237–41.

Bernshtein, Sergei Natanovich (1926) Sur l'extension du théorème limite du calcul des probabilités aux sommes des quantités dépendantes. *Mathematische Annalen*. 97. pp. 1–59.

Bienaymé, Irenée Jules (1853) Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrées. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences*. 37. pp. 309–24.

Bohlmann, Georg (1901) Lebensversicherungs-Mathematik. In Meyer, W. F. (ed.) *Encyklopädie der Mathematischen Wissenschaften*. I. 2. ID4b. Leipzig: Teubner.

Borel, Émile (1905) Remarques sur certaines questions de probabilité. *Bulletin de la Societé Mathématique de France*. 33. pp. 123–8.

Borel, Émile (1909) Les probabilités denombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo*. 27. pp. 247–71.

Brush, Stephen (1976) *The Kind of Motion We Call Heat*. 2 Vols. Princeton, NJ: University Press.

Cantelli, Francesco (1917) Sulla probabilità come limite della frequenza. *Atti della Reale Accademia dei Lincei, Rendiconti.* 26. pp. 39–45.
Cauchy, Augustin Louis (1853) Mémoire sur les résultats moyens d'un très-grand nombre des observations. *Comptes rendus hebdomadaires des séances de l'Académie des Sciences.* 37. pp. 381–5.
Chebyshev, Pafnutii Lvovich (1846) Démonstration élémentaire d'une proposition générale de la théorie des probabilités. *Journal für die reine und angewandte Mathematik.* 33. pp. 259–67.
Chebyshev, Pafnutii Lvovich (1867) Des valeurs moyennes. *Journal de mathématiques pures et appliquées*. 2. 12. pp. 177–84.
Chebyshev, Pafnutii Lvovich (1874) Sur les valeurs limites des integrals. *Journal de mathématiques pures et appliquées*. 2. 19. pp. 157–60.
Chebyshev, Pafnutii Lvovich (1887/1890) Sur deux théorèmes relatifs aux probabilités. *Acta mathematica.* 14. pp. 305–15. (Originally published in Russian.)
Cramér, Harald (1937) *Random Variables and Probability Distributions.* Cambridge: Cambridge University Press.
Doob, Joseph Leo (1953) *Stochastic Processes.* New York, NY: Wiley.
Daston, Lorraine (1988) *Classical Probability in the Enlightenment*. New Jersey, NJ: Princeton University Press.
Ehrenfest, Paul & Tatjana (1911/1959) *The Conceptual Foundations of the Statistical Approach in Mechanics.* Ithaca, NY: Cornell University Press. (Originally published in German.)
Einstein, Albert (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten supendierten Teilchen. *Annalen der Physik*. 4. 17. pp. 549–60.
Feller, Willy (1935) Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift.* 40. pp. 521–59.
Feller, Willy (1937) Über das Gesetz der großen Zahlen. *Acta litterarum ac scientiarum regiae universitatis hungaricae Francisco-Iosephinae, sectio scientiarum mathematicarum*. 8. pp. 191–201.
Fischer, Hans (2011) *A History of the Central Limit Theorem.* New York, NY: Springer.
Fréchet, Maurice (1930) Sur la convergence en probabilité. *Metron.* 8. pp. 1–48.
Fréchet, Maurice (1936/38) *Recherches théoriques modernes sur le calcul des probabilités*. 2 Vols. Paris: Gauthier-Villars.
Gauss, Carl Friedrich (1809/1906) *Theoria motus corporum coelestium*. Hamburg: Perthes & Besser. (Reprinted in *Werke*. Bd. 7. pp. 3–280. Leipzig: Teubner.)
Gauss, Carl Friedrich (1823/1880) Theoria combinationis observationum erroribus minimis obnoxiae, pars prior. *Commentationes societatis Regiae scientiarum Goettingensis recentiores*. 5. (Reprinted in *Werke*. Bd. 4. pp. 3–26. Göttingen.)
Gnedenko, Boris Vladimirovich and Kolmogorov, Andrei Nikolaevich (1949/1954) *Limit Distributions for Sums of Independent Random Variables.* Boston, MA: Addison-Wesley. (Originally published in Russian.)
Hald, Anders (1998) *A History of Mathematical Statistics from 1750 to 1930*. New York, NY: Wiley.
Hilbert, David (1900) Mathematische Probleme, Vortrag, gehalten auf dem internationalen Mathematiker Kongress zu Paris 1900. *Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, math.-phys. Klasse*. 3. pp. 253–97.
Hochkirchen, Thomas (1999) *Die Axiomatisierung der Wahrscheinlichkeitsrechnung und ihre Kontexte*. Göttingen: Vandenhoeck & Ruprecht.
Khinchin, Aleksandr Yakovlevich (1924) Über einen Satz der Wahrscheinlichkeitsrechnung. *Fundamenta mathematicae.* 6. pp. 9–20.

Khinchin, Aleksandr Yakovlevich (1933) *Asymptotische Gesetze der Wahrscheinlichkeitsrechnung.* Berlin: Springer.

Kolmogorov, Andrei Nikolaevich (1928) Über die Summen durch den Zufall bestimmter unabhängiger Größen. *Mathematische Annalen.* 99. pp. 309–19.

Kolmogorov, Andrei Nikolaevich (1931) Über die analytischen Methoden in der Wahrscheinichkeitsrechnung. *Mathematische Annalen.* 104. pp. 415–58.

Kolmogorov, A. (1933/1950) *Foundations of the Theory of Probability.* New York, NY: Chelsea Publ. (Originally published in German under the title *Grundbegriffe der Wahrscheinlichkeitsrechnung*.)

Laplace, Pierre-Simon (1774) Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie Royale des Sciences de Paris.* 6. pp. 621–56. (Reprinted (1891) in *Œuvres completes*. Vol. VIII. pp. 27–65, Paris: Gauthier-Villars.)

Laplace, Pierre-Simon (1812) *Théorie analytique des probabilités.* Paris: Ve. Courcier. (2nd ed. 1814, 3rd ed. 1820. Reprint of 3rd ed. in (1886) *Œuvres completes*. Vol. VII. Paris: Gauthier-Villars.)

Lévy, Paul (1922) Sur la détermination des lois de probabilité par leurs fonctions caractéristiques. *Comptes rendus hebdomadaires de l'Académie des Sciences de Paris.* 175. pp. 854–56.

Lévy, Paul (1935) Propriétés asymptotiques des sommes des variables aléatoires indépendantes ou enchainées. *Journal de mathématiques pures et appliquées.* 14. pp. 347–402.

Lévy, Paul (1937) *Théorie de l'addition des variables aléatoires*. Paris: Gauthier-Villars.

Lindeberg, Jarl Waldemar (1922) Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift.* 15. pp. 211–25.

Lyapunov, Aleksandr Mikhailovich (1901) Nouvelle forme du théorème du calcul des probabilités. *Mémoires de l'Académie Impériale des Sciences de St.-Petersbourg. VIII^{e} Série, Classe Physico-Mathématique*. 12. pp. 175–91.

Markov, Andrei Andreevich (1899/2004) The Law of Large Numbers and the Method of Least Squares. In Sheynin, O. B. (ed.), *Probability and Statistics, Russian Papers*. pp. 130–42. Berlin: NG Verlag. (Originally published in Russian.)

Markov, Andrei Andreevich (1907/2004) The Extension of the Law of Large Numbers onto Quantities Depending on Each Other. In Sheynin, O. B. (ed.), *Probability and Statistics, Russian Papers*. pp. 143–58. Berlin: NG Verlag. (First published in Russian.)

Maxwell, James Clerk (1860/1890) Illustrations of the Dynamical Theory of Gases. *Philosophical Magazine.* 19, 20. (Reprinted in *Scientific Papers*. Vol. 1. pp. 377–409. Cambridge: Cambridge University Press.)

Maxwell, James Clerk (1879/1890) On Boltzmann's Theorem on the Average Distribution of Energy on a System of Material Points. *Transactions of the Cambridge Philosophical Society* 12. (Reprinted in *Scientific Papers*. Vol. 2. pp. 713–41. Cambridge: Cambridge University Press.)

Mehrtens, Herbert (1990) *Moderne-Sprache-Mathematik.* Frankfurt: Suhrkamp.

Shafer, Glenn and Vovk, Vladimir (2006) The Sources of Kolmogorov's "Grundbegriffe". *Statistical Science.* 21. pp. 70–98.

Von Mises, Richard (1919a) Fundamentalsätze der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift.* 4. pp. 1–97.

Von Mises, Richard (1919b) Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift.* 5. pp. 52–99.

Von Plato, Jan (1994) *Creating Modern Probability.* Cambridge: Cambridge University Press.

Wald, Abraham (1937) Die Widerspruchsfreiheit des Kollektivbegriffes der Wahrscheinlichkeitsrechnung. *Ergebnisse eines mathematischen Kolloquiums*. Heft 8. pp. 38–72.

CHAPTER 6

THE ORIGINS OF MODERN STATISTICS

The English Statistical School

JOHN ALDRICH

6.1 Introduction

"Modern statistical theory originated in England, and is today advancing faster there than in any other country." So Harold Hotelling (1930: p. 186) informed the American Statistical Association in 1930: the originator was Karl Pearson and Ronald Fisher was chiefly responsible for the advances of the time. Both traced their work back to Francis Galton, "that versatile and somewhat eccentric man of genius" as Fisher (1956: p. 1) once described him. Galton's obituarist (Anon 1911: p. 315) described his achievement, "As an incidental result of his statistical researches in biology, he was the parent of modern statistical methods."

Francis Galton (1822–1911), Karl Pearson (1857–1936), and Ronald Aylmer Fisher (1890–1962) are the chief figures of the English statistical school. This chapter traces the origins and consolidation of the school, from the 1860s to the 1930s, with the period before 1890 belonging to Galton, 1890–1920 to Pearson and the period after 1920 to Fisher. The chapter concludes with the school's eclipse: in 1930 English influence was at a peak but ten—certainly twenty—years later Hotelling would have judged that statistical theory was advancing faster in his own United States.

Hotelling and Galton's obituarist—most probably Yule—knew there was earlier work, though how much has only been revealed by modern historians: cf. Anders Hald's *History of Mathematical Statistics from 1750 to 1930* and Stephen Stigler's *The History of Statistics: The Measurement of Uncertainty before 1900*. The English development was distinctive in being associated with biology and became more so by making a speciality of certain topics, such as correlation. Such was the continuity of purpose and sense of self-sufficiency evinced that it is natural to speak of an English *school*. Laplace, Gauss, and Quetelet were acknowledged, in varying degrees, as intellectual ancestors, but the English statisticians took little from

their foreign contemporaries. Yet the school was fractured, for Pearson, and Fisher agreed on little beyond an admiration for Galton and on the importance of their own work.

There is an extensive literature on the English statistical school: the Galton–Pearson phase is very well covered in Stigler's *History*, Hald's *History* brings out connections between the mathematical theory of the English school and that of the Continental statisticians, while MacKenzie (1981) gives a sociologist's account of the school. There are biographies of the main figures including Bulmer (2003) and Gillham (2002) on Galton, E. S. Pearson (1936/8) and Porter (2004) on Pearson and Box (1978) on Fisher; short biographies of other figures can be found in Heyde and Seneta (2001). There is a substantial journal literature and there are bibliographies for Pearson and Fisher by Aldrich (2013 and 2012). The work of Galton, Pearson, and Fisher is as much part of the history of biology and is treated in histories of genetics and evolution, such as Provine (1971) and Gayon (1998).

6.2 Statistics as a Study

In 1834 English enthusiasts for statistics came together to found the Statistical Society of London. Its mission, as stated in the first volume of the *Journal*, was "the collection and comparison of Facts which illustrate the condition of mankind, and tend to develop the principles by which the progress of society is determined." (Anon 1838: p. 1). In 1887 the society became the Royal Statistical Society without any change of mission. Galton belonged to the Society and actually believed that his statistical research on heredity held the key to the "progress of society", but he never presented his results to the Society—biology was not their concern.

In 1920 Karl Pearson set the failings of the older statisticians against the ideal of statistics to which he had already devoted 25 years (quoted by E. S. Pearson 1938: p. 164):

> The object of [the English school of mathematical statistics] was to make statistics a branch of applied mathematics with a technique and nomenclature of its own, to train statisticians as men of science, to extend, discard or justify the meagre processes of the older school of political and social statisticians, and in general to convert statistics in this country from being the playing field of *dilettanti* and controversialists into a serious branch of science, which no man could attempt to use effectively without adequate training, any more than he could attempt to use the differential calculus, being ignorant of mathematics.

The "older school" was concerned mainly with economics and vital statistics and only a few economists and demographers were interested in using mathematical methods based on probability theory. The most notable was F. Y. Edgeworth (1845–1926) who began contributing mathematical papers to the Society's *Journal* in 1883. Edgeworth began as an older counsellor to Pearson, became a rival and ended invisible. Edgeworth had only one real follower, the economic statistician Arthur Bowley (1869–1957). In fact the non-economist Pearson had much more influence on economists.

Pearson never joined the Society but many of his followers did, carrying with them the conception of "a branch of applied mathematics" though not the biological mission it originally served. Slowly the Society changed—Aldrich (2008) charts the process—and in 1933–4 an Industrial and Agricultural Research Section and *Supplement* to the *Journal of*

the Royal Statistical Society were set up to cater for the new interests and the mathematical level. Twenty years later Fisher (1956: p. 1)—for whom statistics had always been a branch of applied mathematics—could report that statistics served diverse "technological, commercial, educational and administrative purposes", as well as a "scientific field" of indefinite extent.

The professionalism desired by Pearson is reflected in the circumstances of the big three. Galton was too serious to be a *dilettante* but he was a private scholar—rich and very well connected—who influenced others through his writing, participation in learned bodies such as the British Association and the Royal Society, and money. Karl Pearson and Ronald Fisher were professional scientists who ran laboratories, trained students, and edited journals.

6.3 Francis Galton

Galton (1883: p. 33) once gave as the object of statistical science "to discover methods of condensing information concerning large groups of allied facts into brief and compendious expressions suitable for discussion." Galton devised methods of condensation, both graphical and numerical; the ogive and percentiles had some take-up and they formed the basis for Elderton and Elderton's (1909) *Primer of Statistics*. However his innovations in descriptive statistics did not make Galton the "parent of modern statistical methods"—that distinction was a by-product of his investigations into heredity and anthropometry.

Unlike writers on the theory of errors and unlike Pearson and Fisher, Galton did not use probability to make statistical inferences. He *did* use probability to model phenomena and this attitude and the models he devised were his contribution to the English statistical school; his successors provided the relevant probabilistic statistical inference. Galton's researches into heredity were spread over nearly 50 years and their many twists and turns are followed by Bulmer (2003); for an account focusing on Galton's contribution to statistics see Stigler (1986: chapter 8).

Galton's crucial contributions came later but his first large study of inheritance, *Hereditary Genius* (1869), set the tone for the entire research programme. Galton (1869: p. 1) expected research into hereditary genius, or inherited ability, to bring an increase of knowledge and important practical benefits:

> I propose to show in this book that a man's natural abilities are derived by inheritance, under exactly the same limitations as are the form and physical features of the whole organic world. Consequently […] it would be quite practicable to produce a highly-gifted race of men by judicious marriages during several consecutive generations.

While Galton had the application to eugenics in mind from the beginning, he only gave his full attention to the topic 30 years later.

Galton (1869: p. vi) described what he had achieved in the following terms:

> I may claim to be the first to treat [the theory of hereditary genius] in a statistical manner, to arrive at numerical results, and to introduce the "law of deviation from an average" into discussions of heredity.

The book is almost entirely taken up with the results of research into the family connections of the eminent in fields ranging from the law to competitive rowing. Information from handbooks is condensed into tables and conclusions drawn from them. Galton had learnt of the "law of deviation from an average" or normal distribution from Quetelet, who had applied it to anthropometric data, though not to heredity. The distribution did not play any part in the book's genealogical researches but it would figure prominently later.

The most detailed of the book's case studies treats the judges of England from 1660 to 1865. There is a lot of condensing into tables and informal testing of the theory of hereditary genius. At one point Galton (1869: p. 67) carries out a kind of significance test when he compares the numbers of eminent relations with what would be expected if "natural gifts were due to mere accident, unconnected with parentage".

> If it be a hundred to one against some member of the family, within given limits of kinship, drawing a lottery prize, it would be a million to one against three members of the same family doing so (nearly, but not exactly, because the size of the family is limited), and a million millions to one against six members of the family doing so.

On the contrary, Galton finds

> there are nearly as many cases of two or three eminent relations as of one eminent relation […] It is therefore clear that ability is not haphazard but that it clings to certain families.

While Galton continued to be interested in the genealogy of the eminent, the book's speculative final chapter has a closer relationship to the works that would be seminal for the English statistical school. The "General Considerations" introduce some speculations on heredity linked to the theory of pangenesis in Darwin's *Variation of Animals and Plants under Domestication*. The spirit of much of Galton's work is captured in his (1869: p. 371) declaration:

> I do not see why any serious difficulty should stand in the way of mathematicians, in framing a compact formula, based on the theory of Pangenesis, to express the composition of organic beings in terms of their inherited and individual peculiarities, and to give us, after certain constants had been determined, the means of foretelling the average distribution of characteristics among a large population of offspring whose parentage was known.

Galton sketches a mathematical model of inheritance although, unlike his later efforts, this had no probabilistic element.

Through the 1870s Galton continued to work on inheritance, writing theoretical papers and trying to establish experimentally whether Darwin's theory could be true. The papers contained a certain amount of probabilistic reasoning, but it was not carried through. Thus in "On Blood Relationship" Galton (1871–2: p. 400) proposes an urn metaphor:

> An approximate notion of the nearest conceivable relationship between a parent and his child may be gained by supposing an urn containing a great number of balls, marked in various ways, and a handful of them to be drawn out at random as a sample; this sample would represent the person of a parent. Let us next suppose the sample to be examined, and a few handfuls of new balls to be marked according to the patterns of those found in the sample, and to be thrown along with them back into the urn. Now let the contents of another urn, representing the influences of the other parent, be mixed with those of the first. Lastly suppose a second sample to be drawn out of the combined contents of the two urns, to represent the offspring.

Galton did nothing with this urn scheme. In 1875 however he brought to fruition an effective piece of modelling when he induced Henry Watson to analyse a branching process for the "extinction of families."

With "Typical Laws of Heredity" (1877) Galton's mathematical modelling took on a new seriousness: this was a genuine application of the "law of deviation from an average" to the subject of heredity. The central concept was "reversion", a biological concept that Galton formalized as a stable first-order autoregressive process with a generation as the unit of "time. " For the formal development Galton used properties of the normal that he found in Airy's theory of errors textbook. Unlike "Blood Relationship" this was not a wholly theoretical paper, for Galton investigated reversion in data on sweet peas. The degree of reversion was estimated from the data, but no use was made of the estimation theory in Airy's book.

The concept of reversion (later called "regression") would become part of English statistics. That such a phenomenon existed in biological populations belonged to a body of knowledge about inheritance that Galton built up. The best known of these biological facts was "the law of ancestral heredity." The law would go through numerous reformulations but Galton (1886: p. 261) first expressed it so:

> The influence, pure and simple, of the mid-parent may be taken as 1/2, of the mid-grandparent 1/4, of the mid-great-grandparent 1/8, and so on. That of the individual parent would therefore be 1/4, of the individual grandparent 1/16 of an individual in the next generation 1/64, and so on.

A concept with a close mathematical relationship to regression was correlation and this would be Galton's most influential contribution to statistical methods. He introduced the concept in his "Co-relations and their measurement, chiefly from anthropometric data." The anthropometric data was on the size of limbs, either of the same person or of relatives. Galton (1888: p. 135) defined correlation and explained how it arose:

> Two variable organs are said to be co-related when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. […] It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes.

As with regression, Galton improvised a way of estimating the degree of correlation. Edgeworth made some progress in relating correlation to known results in probability but the definitive performance came from Pearson in 1896—see below.

Galton's *Natural Inheritance* (1889) summarized his work on heredity, but though he continued to write for another twenty years, his most influential work on statistics was done. He was remembered in the *Statistical Journal* as "the parent of modern statistical methods." In the Preface to volume 1 of the vast *Life, Letters and Labours of Francis Galton* Karl Pearson (1914: p. vii) emphasized his importance for biology, comparing him to his half-cousin Charles Darwin:

> Darwinism needs the complement of Galtonian method before it can become a demonstrable truth; it requires to be supplemented by Galtonian enthusiasm before it can exercise a substantial influence on the conscious direction of race evolution.

Pearson had taken over management of the "Galtonian method", while for a decade Galton and Pearson had been lavishing "Galtonian enthusiasm" on "the conscious direction of race evolution" or eugenics; Mazumdar (1992) recounts the history of eugenics in Britain and the part Galton, Pearson, and Fisher played in it.

6.4 Karl Pearson

No disciple ever paid a more fulsome tribute than the Galton *Life*, and yet Pearson was not a disciple in any simple sense. He was an established applied mathematician (theoretical physicist), a professor at University College London, when a colleague, the zoologist W. F. R. Weldon (1860–1906), brought him into the field they later called biometry. Weldon *was* influenced by Galton and had been applying Galton's statistical methods, including correlation, in the light of his belief that "the problem of animal evolution is essentially a statistical problem." Pearson joined in, developing new techniques and eventually a new theory of statistics. Pearson's first work was not influenced by Galton, and he was slow to become a disciple.

Pearson had a wide range of social, philosophical, literary, and historical interests, and two early projects brought together his philosophical and scientific concerns. He prepared a manuscript left by W. K. Clifford for publication, writing about a third of the final text himself. *The Common Sense of the Exact Sciences* set out to explain the basic principles of mathematics in a non-technical way and Pearson (1888) added his own ideas for the reform of mechanics—ideas that were close to those of Ernst Mach. Pearson expounded his general philosophy of science in *The Grammar of Science* (1892). This depicts the scientific method as "the orderly classification of facts followed by the recognition of their relationship and recurring sequences." One of the chapters is "Cause and Effect. Probability" and the following passage (1892: p. 136) explains the connection:

> That a certain sequence has occurred and recurred in the past is a matter of experience to which we give expression in the concept causation; that it will continue to recur in the future is a matter of belief to which we give expression in the concept probability. Science in no case can demonstrate any inherent necessity in a sequence, nor prove with absolute certainty that it must be repeated.

However, the probability of a repetition can be evaluated and Pearson (1892: pp. 168ff.) proposes a modification of Laplace's law of succession. Pearson had attended lectures on probability and the theory of errors at Cambridge, but this was the first time he used the knowledge he had imbibed.

In the next ten years Pearson made his most important contributions to statistics: the method of moments, the Pearson system of curves, correlation and the chi-squared test. The first two "Contributions to the Mathematical Theory of Evolution" were concerned with fitting a distribution to measurements of a biological population and they used a method of Pearson's own devising—the method of moments. The idea came from mechanics, where the concept of a moment was fundamental, and the calculation of moments was one of the demonstration pieces of graphical statics. In the first "Contribution" Pearson (1894) used the method of moments to dissect an abnormal frequency curve into two normal curves.

It is likely that he first tried the method of "inverse chances" which he had known from his undergraduate days: this finds the maximum of the posterior distribution based on a uniform prior. However, the problem of estimating a mixture of normal distributions does not yield to the method. The method of moments produced an answer, but Pearson had no further justification for the method than that it gave the usual results for the parameters of the normal distribution.

Pearson had no faith in the universality of the normal distribution, nor did he have much confidence in the traditional theory of errors, founded upon normality. He considered "skew variation" the usual condition and the second of the "Contributions" (1895) presents a system of frequency curves which would be of use in "many physical, economic and biological investigations." From probabilistic considerations—which now seem contrived—Pearson obtained a basic differential equation governing the probability density function, or frequency curve, $y = y(x)$:

$$\frac{dy}{dx} = \frac{y(x+a)}{b_0 + b_1 x + b_2 x^2} \tag{6.1}$$

The constants determine the form of the curve. Pearson developed methods for choosing an appropriate curve and used the method of moments for estimating the constants. Tables were produced to ease the burden of calculation.

With the third of the "Contributions," "Regression, heredity and panmixia" (1896), Pearson at last made contact with Galton. He systematized some of the concepts Galton had introduced, including correlation and regression, placing them in the framework of the multivariate normal distribution; later papers would reformulate the law of ancestral heredity. Pearson (1896: p. 259) presented definitions such as this: "Heredity.—Given any organ in a parent and the same or any other organ in its offspring, the mathematical measure of heredity is the correlation of these organs for pairs of parent and offspring." Pearson (1896: p. 292) also offered some methodological reflections on the nature of the biometric project: the statistical formulae "make not the least pretence to explain the mechanism of inheritance. All they attempt is to provide a basis for the quantitative measure of inheritance—a schedule, as it were, for tabulating and appreciating statistics." On this issue Pearson appears to have influenced Galton, who (1908: p. 308) recalled the origins of his work on heredity without mentioning his hypotheses about mechanisms:

> ...it seemed most desirable to obtain data that would throw light on the Average contribution of each Ancestor to the total heritage of the offspring in a mixed population. This is a purely statistical question, the same answer to which would be given on more than one theoretical hypothesis of heredity, whether it be Pangenetic, Mendelian or other.

The 1896 paper's most influential contribution to statistical theory was a new method for estimating correlation that replaced the methods used by Galton and Edgeworth; Stigler (1986: chapters 9 and 10) reviews all this activity. Pearson's reasoning is not entirely clear, but it appears to be based on the method of inverse probability: Pearson (1896: p. 265) chooses the value of the correlation for which "the observed result is the most probable." Pearson also gave a formula for the probable error of r—the conventional measure of accuracy—which was needed for significance tests. The probable error was a measure of dispersion for the normal approximation to the posterior of the correlation.

As yet Pearson had no procedure for calculating probable errors for the values found by his method of moments. Pearson and Filon (1898: pp. 231–6) 'generalized' the procedure used for correlation to obtain probable errors for the method of moments estimators. The generalization was invalid because the procedure was based on the Bayes posterior for the correlation and no Bayes posterior was associated with the method of moments estimator. In Pearson's work of this time it is unclear whether the probable error is a property of a posterior distribution or of a sampling distribution; in his later work it was clearly the latter. However, Pearson continued to produce Bayesian analysis alongside frequentist work—see e.g. Pearson (1920)—and from the later perspective of Fisher, Neyman or Jeffreys, he seems to have had no coherent position on the principles of statistical inference.

The last of Pearson's great contributions was the χ^2 goodness of fit test introduced in 1900. This completed the set of tools for fitting univariate distributions to data. Pearson (1900) was not in the "Contributions" series and it contained no analysis of biological data; it was an exercise in pure statistical technique—even the general analysis of Pearson and Filon had been coupled with biological analysis. Pearson would work for another three decades and write hundreds of articles, but none were so important for general statistical methodology as the five published in 1894–1900. The years 1900–14 brought the institutions that transformed the part-time occupation of a professor of mechanics into an international movement.

The voice of the international movement was *Biometrika*. The journal came out of a crisis in 1900–1, when Pearson had a paper rejected by the *Philosophical Transactions of the Royal Society*. The applied mathematics professor was used to publishing in Series A for papers of a "mathematical or physical character." But in 1900 the latest "Contribution" was sent to William Bateson, a referee for Series B (for papers of a "biological character"), and he rejected it. Pearson told Galton—see K. Pearson (1930: p. 241)—that "if the R.S. people send my papers to Bateson, one cannot hope to get them printed. It is a practical notice to quit. This notice applies not only to my work but to most work on similar statistical lines." Pearson and Weldon, with Galton's support, moral and financial, responded by founding a "Journal for the Statistical Study of Biological Problems"—*Biometrika*. The journal, which would be edited by Pearson until his death in 1936, would be a cornerstone of the English school; see Aldrich (2013b).

The early days of *Biometrika* were overshadowed by worsening relations with William Bateson (1861–1926): following the "rediscovery" of Mendel, Bateson was arguing that the statistical study of heredity had no value. Pearson had always insisted that his statistical formulae were descriptive rather than explanatory but that did not mean they had no bearing on explanatory theories, as he (1904: p. 85) pointed out:

> [The] biometric or statistical theory of heredity does not involve a denial of any physiological theory of heredity, but it serves in itself to confirm or refute such a theory. Mendelian formulae analytically developed for randomly mating populations are either consistent or not with the biometric observations on such populations. If they are consistent, it shows their possibility, but does not prove their necessity. If they are not, it shows they are inadequate.

On examination the gross incompatibility that Weldon and Bateson saw disappeared, yet still Pearson (1904: p. 85) found against Mendelian theory:

> The present investigation shows that in the theory of the pure gamete there is nothing in essential opposition to the broad features of linear regression, skew distribution, the

> geometric law of ancestral correlation, etc., of the biometric description of inheritance in populations. But it does show that the generalised theory here dealt with is not elastic enough to account for the numerical values of the constants of heredity hitherto observed.

However, while meeting the threat posed to biometry by Mendelism was a vital concern for Pearson and Weldon, their followers either were not so interested or were sympathetic to Mendelism.

The early years of the twentieth century brought organizational changes which strengthened the movement and lasted until Pearson's retirement; Magnello (1999) reviews the various Pearson enterprises. In 1903 Pearson established the Biometric Laboratory. This drew visitors from all over the world—from America the biologists Raymond Pearl and Arthur Harris and the economist H. L. Moore. In 1907 Pearson took over responsibility for a research unit founded by Galton which was reconstituted as the Francis Galton Laboratory of National Eugenics. Pearson saw his role in eugenics as providing the scientific foundations, and he addressed other experts rather than the public directly. The Laboratory researched human pedigrees, but it also produced controversial reports on the role of inherited and environmental factors in tuberculosis, alcoholism, and insanity. A bequest from Galton led to the establishment of the Sir Francis Galton Professorship of National Eugenics in 1914 and a Department of Applied Statistics. Pearson no longer had to teach engineering students applied mathematics.

Pearson realized that the methods he had devised for biometry had other uses and he and his co-workers applied them to all manner of subjects. The ideas of two such followers, G. U. Yule and W. S. Gosset, are described in the next section.

6.5 Pearson's Early Students

Galton enlisted the help of several talented mathematicians including Henry Watson, George Darwin, Donald MacAlister, and Hamilton Dickson, but a single job for him did not turn them into statisticians. Pearson, by contrast, was a professional teacher: students came to him as undergraduates and older visitors came to learn how to apply his methods to problems of their own. To illustrate the experience of working with Pearson, consider a student, Yule, and a visitor, Gosset, both of whom became important figures in statistics.

G. Udny Yule (1871–1951) studied with Pearson as an undergraduate and went on to become his assistant. Yule was in Pearson's first class on statistical theory in 1893 and was soon contributing to the theory of correlation and regression; after 1900 he developed a parallel theory of association for attributes. With Pearson's blessing Yule applied the new statistical techniques to social problems and established himself in the Statistical Society. Yule introduced a second interpretation of correlation based, not on Galton's notion of common causes, but on the idea that one correlated variable is partial cause of another: the conception forms the basis for his investigation of pauperism, a classic Statistical Society topic. Yule's "An Investigation into the Causes of Changes in Pauperism in England" (1899) took regression away from biology and from the multivariate normal distribution. Yule also provided careful discussions of illusory correlations; see Aldrich (1995).

Galton and Pearson were virtuosi prepared to take on subjects outside their chief field of heredity, but Yule might be judged the first modern statistician—someone with a specialized type of competence but ready to work in any field. For Yule that might be economics, medicine, accidents, or social policy. He did some early work on population genetics and produced the first examination of the consequences for population outcomes of Mendelian principles. Yule (1902: p. 236) concluded by declaring:

> It is, however, essential, if progress is to be made, that biologists—statistical or otherwise—should recognise that Mendel's Laws and the Law of Ancestral Heredity are not necessarily contradictory statements, one or other of which must be mythical in character, but are perfectly consistent the one with the other and may quite well form parts of one homogeneous theory of heredity.

Yule disliked Bateson's polemics but he strove for an accommodation between the two systems of ideas.

Around 1905 Yule broke with Pearson, partly, it seems, because of disagreements over the analysis of association and partly over Mendelian theory. After Pearson's death Yule (1936: p. 101) reflected on the separation, "Those who left him and began to think for themselves were apt, as happened painfully in more instances than one, to find that after a divergence of opinion the maintenance of friendly relations became difficult, after express criticism impossible." Yule found a new intellectual home in the Statistical Society—though when he wrote the obituary of Galton he was one of the very few members of the Society using "modern statistical methods"—and he taught in the School of Agriculture at Cambridge for nearly 20 years. However his main influence was probably through his *Introduction to the Theory of Statistics* (1911). The book emphasized relations between variables—correlation and association—the topics in which Yule made his biggest contribution. The Pearson system of curves and the associated method of moments are not treated, although they are in the more authentically Pearsonian textbook W. P. Elderton wrote for his fellow actuaries, *Frequency-Curves and Correlation* (1907).

William Sealy Gosset (1876–1937) was not an academic, but a chemist who worked for Guinness, the brewer. Gosset saw the potential for using the theory of errors in the work of the brewery and taught himself the subject from textbooks. The application was not straightforward and Gosset went to study for a year with Karl Pearson to improve his knowledge. With "Student" as a pen-name, Gosset started to publish his results in *Biometrika*, including two notable papers in 1908, one (1908a) on the mean of the normal distribution, an investigation in the theory of errors using Pearsonian methods, and the other (1980b) on the Pearsonian topic of distribution of the correlation coefficient. Both studies attempted to replace large-sample results—of the kind Pearson had produced—with exact results. Gosset's work for Guinness and the farms that supplied it led to work on agricultural experiments, though he published nothing on this subject in *Biometrika*. Gosset was not very interested in biometry and the biometricians were not very interested in what he did; the normal mean problem belonged to the theory of errors which Pearson considered outdated. Gosset was a marginal figure in the English statistical school until Fisher built on his small-sample work and made his name one of the most celebrated in English statistics. Unlike Yule (and Fisher), Gosset always enjoyed good relations with Pearson and published all his important work in *Biometrika*.

In 1933 Pearson retired from his positions at University College but he went on writing and editing *Biometrika* until his death. However, the initiative in mathematical statistics had long since passed to Ronald Fisher. Before the First World War Pearson had been a great controversialist, but he largely ignored Fisher. He wrote only one piece criticizing Fisher, the posthumous article, Pearson (1936), which defended the method of moments against Fisher's criticism.

6.6 Ronald Fisher

In the course of the 1920s R. A. Fisher replaced Pearson as leader of the biometric school and of English mathematical statistics generally; see the biography by Box (1978) and the bibliography by Aldrich (2012). Fisher found solutions for some of the outstanding problems of the Pearson programme when he derived the exact distribution of the correlation coefficient in 1915, of the partial correlation coefficient in 1921, and of the multiple correlation in 1928. He noisily rejected some Pearsonian methods, including the method of moments, and quietly dropped some topics, such as the system of curves and the quest for generalizations of correlation and regression to skew distributions. He was responsible for some major reorientations: statistical inference became an imposing intellectual structure rather than a box of tools serving applied statistics, and the attention of mathematical statisticians was turned from the analysis of observational data towards the design and analysis of experiments. While Fisher attacked Pearson's practice at several points, it was not an instance of one system confronting another, for there was no Pearsonian system. It was Fisher who found systematic things to say about the method of moments and inverse probability—systematically damning things.

Fisher's future was prefigured in two papers he wrote as an undergraduate. His first published paper, "On an Absolute Criterion for Fitting Frequency Curves" (1912), proposed a method (the later maximum likelihood) arguing that, as an "absolute criterion" (invariant to parameter transformation), it was superior to the method of moments and to inverse probability. The other paper departed from Pearson's position on biology. In an address to an undergraduate society Fisher envisaged a synthesis of Mendelism and Biometry. Fisher was as committed to eugenics as Pearson but he belonged to a branch of the movement of which Pearson did not approve; see Mazumdar (1992).

After graduating Fisher did not settle to a career and was a schoolmaster when he published his "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population" (1915) and "The Correlation between Relatives on the Supposition of Mendelian Inheritance" (1918). The first solved the problem proposed by "Student" (1908b) and marked a new era in the exact theory of sampling distributions. The second vindicated the promise of "Mendelism and Biometry" by showing how Pearson's biometric results could be explained by Mendelian theory.

Fisher's career took off in 1919 when he was given a temporary job as a statistician at Rothamsted Experimental Station; the position was soon made permanent and a statistics department established. Fisher's first task was to analyze the data on crop yields that Rothamsted had accumulated over decades. Fisher's methods were adaptations of methods from the theory of errors, notably the analysis of variance and a reformulation of regression

that identified it with the traditional theory of errors structure. Fisher then went from analyzing experiments to designing them. There had been some earlier work but Fisher raised the subject to a new level. His principles of randomization, replication, and blocking were presented in *Statistical Methods for Research Workers* (1925) and more fully in *The Design of Experiments* (1935).

Fisher's early years at Rothamsted were as productive as the '90s had been for Pearson: as well as striking out in new directions, Fisher continued to work on statistical and genetical theory. Fisher (1921) combined advocacy of what has been called the likelihood approach and an attack on the Bayesian approach. Pearson rejected Fisher's paper for *Biometrika*, which led to the permanent estrangement of the two; see Aldrich (1997) for an account of the paper and the estrangement. Fisher's next piece on estimation, the monumental "On the Mathematical Foundations of Theoretical Statistics" (1922), was probably the most intellectually ambitious work of the English school; see Stigler (2005) for an appreciation.

Fisher (1922: p. 311) describes the statistician's task in its "most concrete form" as the "reduction of data":

> A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

The language is reminiscent of Galton's statistical science as "condensation" but Fisher's statistician constructs a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. Pearson's statistician did the same, but in Fisher's formulation "information" is formalized and the concept of "sufficiency" expresses the idea of containing "the whole of the relevant information."

In Fisher's paper "theoretical statistics" meant estimation and maximum likelihood was presented as an efficient way of extracting information from the data—information and efficiency being new technical concepts. Looking back, one part was concerned with correcting the results of Pearson and Filon (1898); looking forward, establishing—or disproving—the claims of the paper occupied statistical theorists for decades—see Stigler (2007) for an account.

Fisher also reconstructed the theory of Pearson's chi-squared test and extended the scope of "Student"'s (1908a) distribution. These developments, like the analysis of variance, relied on a new system of distribution theory, based on the interrelation of the normal, χ^2 t, and z (a function of the modern F) distributions. This system formed the basis of Fisher's *Statistical Methods for Research Workers* (1925) a manual instructing researchers (primarily of biology) in the methods based on this system. While the book contained some material on estimation, it mainly provided instruction in significance testing. The book and the tables it contained revolutionized applied statistics, replacing the methods Pearson had introduced at the turn of the century.

Fisher's genetical research at Rothamsted concentrated on evolution, on integrating Mendelian theory with Darwin's theory of natural selection. His ideas on evolution were brought together in his book *Genetical Theory of Natural Selection* (1930a). He argued that Mendelism with its view of particulate inheritance did not contradict Darwinism but was consistent with it. With Sewall Wright and J. B. S. Haldane, Fisher is generally recognized as one of the architects of "The Modern Synthesis." Although Fisher made more of a

contribution to the "mathematical theory of evolution" than did Pearson, his contributions to statistics were less entangled with biology and he contributed to the movement of statistics away from biology.

In 1930 Fisher's inference theory took a very surprising turn when he introduced the fiducial argument in his paper "Inverse Probability" (1930b). The method promised to produce post-data probability statements about parameter values without invoking a prior distribution; of course Fisher was opposed to any introduction of a prior distribution. In his lifetime the fiducial argument would prove to be the most disputed of his contributions and after his death its validity was still debated, but with diminished interest and intensity.

Fisher had made Rothamsted a centre for mathematical statistics, but in 1933 he moved to the original one at University College, to replace Pearson as Galton Professor of Eugenics and head of the Galton Laboratory. Fisher was Pearson's natural successor in both statistics and eugenics, but he did not inherit the whole empire, for Pearson's son, Egon Sharpe Pearson (1895–1980) took over the Department of Applied Statistics and became editor of *Biometrika*. This structure did not make for harmony and relations between Fisher and members of Pearson's department, especially its leading theorist, Jerzy Neyman, gradually deteriorated.

In his *History of Mathematical Statistics* Anders Hald (1998: p. 738) writes that R. A. Fisher "almost single-handedly created the foundations for modern statistical science." Hald was referring to its intellectual foundations, especially those for estimation and distribution theory, for as a creator of institutions Fisher achieved much less than Karl Pearson. Fisher created an enduring research group at Rothamsted, but an agricultural research station offers limited scope for education, and Fisher created no *Biometrika*. At University College and in Cambridge (from 1943) Fisher's professorial duties were in genetics, and not in statistics.

6.7 A New Generation

After the First World War Pearson was no longer producing influential new ideas but University College went on attracting and nurturing talent. The most significant newcomer was his son, Egon, who would succeed Karl in several of his roles. Egon Sharpe Pearson (1895–1980) joined after war service and studying mathematics in Cambridge. At first he worked on his father's projects, but he soon came under the influence of Fisher and "Student." Egon was a statistician rather than a biometrician and his chosen field of application was to industrial quality control, a subject created by the American Walter Shewhart. Egon further established a separate identity in his work with the Pole Jerzy Neyman (1894–1981), who came as a visitor to Pearson's laboratory in 1925. Neyman's mathematical background was unlike that of the English mathematical statisticians: he had been educated in the tradition of Russian probability theory and he maintained a strong interest in pure mathematics. The high point of the Neyman and Pearson collaboration was their 1933 paper which presented a general theory of testing, founded on such concepts as Types I and II errors and the fundamental lemma. Their aim was to provide a foundation for the test procedures devised by Karl Pearson, "Student" and Fisher—the ones in Fisher's *Statistical Methods*. Neyman's (1934) "confidence intervals" appeared to be very similar to

Fisher's fiducial intervals, but it gradually emerged that they were not. At first Neyman had good relations with Fisher but these began to deteriorate after Neyman moved from Poland in 1934 to join Egon Pearson's Department of Applied Statistics. In 1938 Neyman moved to Berkeley.

By the end of the '20s Rothamsted had also become a training ground for statisticians. Fisher employed Oscar Irwin, John Wishart and Frank Yates as assistants, while Hotelling came as a visitor. Wishart left for Cambridge to replace Yule in the School of Agriculture, but Wishart also taught mathematics students. The conception of statistics as a branch of applied mathematics was at last taking root at Cambridge, the great centre for mathematics in Britain.

6.8 Harold Jeffreys

Harold Jeffreys (1891–1989) may be seen as a dissenting member of the English statistical school—or as a critical commentator. Like Pearson and Fisher, Jeffreys was trained as a Cambridge applied mathematician but he remained one, specializing in astronomy and geophysics. Between 1919 and 1923 Jeffreys published a series of papers with Dorothy Wrinch. Their approach, which would now be described as Bayesian, reflected the ideas of the Cambridge philosophers W. E. Johnson and C. D. Broad. Like Pearson and Fisher, Jeffreys had been taught the theory of errors as an undergraduate but around 1930 he started to devise new methods and to reconstruct the old in accordance with his theory of probability. Jeffreys then learnt of Fisher's statistical work and adopted some of his concepts and terminology, such as likelihood and sufficiency. Jeffreys's large and idiosyncratically titled *Theory of Probability* (1939) contains a reworking of the statistics of Fisher and Pearson based on the principles of inverse probability. For his Bayesian outlook Jeffreys acknowledged a debt to Pearson's *Grammar of Science*, though Jeffreys was perplexed to find that the principle that supported all his work on inference did not support much of Pearson's. Although Jeffreys was respected as a scientist, his ideas on statistics were rejected by Fisher and most English statisticians. Nor did his ideas get much acceptance from physical scientists.

6.9 A British-American School and Then ...

The Second World War had a profound impact on the English school. There was growth as young mathematicians were recruited to do statistical war work and many stayed to make a career of statistics. It was the same in the United States but on a larger scale and the English school went into relative decline. Some British—and Indian—statisticians moved to the United States but ideas began flowing in the opposite direction.

In the 1940s the *Annals of Mathematical Statistics*, the journal of the Institute of Mathematical Statistics, replaced *Biometrika* as the leading statistical theory journal. In 1939 the *Annals* had a new editorial team which included Samuel Wilks, Hotelling, and

Neyman. All had spent time in England and their research reflected English priorities, so it is possible to speak of a British–American school. However, American mathematical statistics would develop a distinct character. It was tied to probability theory and the *Annals* differed from the English journals in being a journal for probability as well as for statistical theory. English statistics, like English applied mathematics generally, produced some impressive probability—Fisher did outstanding work on stochastic processes on diffusion processes and branching processes—but there was no interest in the foundations of probability.

1939 saw the *Annals* debut of Abraham Wald (1902–50) with an article on statistical decision theory, a search for foundations beneath the foundations. Wald was an applied mathematician of a different kind from that of Fisher or Karl Pearson. He had begun in Vienna as a pure mathematician but then worked in probability, economic theory, and economic statistics to make a living. In 1938 Wald moved to America, where Hotelling introduced him to modern statistical theory, making him in a way one of Fisher's intellectual grandchildren. Wald's mathematics was much more pure than Pearson's or Fisher's and his interpretation of statistical inference as a form of decision-making was remote from theirs.

6.10 Afterwords

Karl Pearson recounted the origins of the English school in great style and at prodigious length in his *Life, Letters and Labours of Francis Galton* (1914–30). Fisher's *Statistical Methods and Scientific Inference* (1956) has a brief, personal and rather sad history of the school. In the Foreword Fisher praised Galton but found nothing to admire in Pearson's intellectual efforts, though he conceded that Pearson's activities "have a real place in the history of a wider movement." There was a past Fisher attended to more closely—the work of Bayes, Boole, and Laplace reviewed in the chapter on "the early attempts and their difficulties." Fisher had long been familiar with these names but this was his first real examination of their ideas—he was inventing a context for his own ideas on statistical inference. He found nothing to praise in his American successors, for he was profoundly unsympathetic to the mathematical statistics that Neyman and Wald were developing in America. The chapter on "some misapprehensions about tests of significance" argues that the Neyman–Wald account of testing is based on a false analogy between "tests of significance" and "acceptance decisions." Tests of significance were originally means by which the "the research worker gains a better understanding of his experimental material." However, the tests have been reinterpreted "as means to making decisions in an acceptance procedure." Fisher's dismay at the work of his successors had a precedent: Pearson (1936) had been dismayed by his work.

References

Airy, G. B. (1861) *On the Algebraical and Numerical Theory of Errors of Observations and the Combination of Observations*. London: Macmillan.

Aldrich, J. (1995) Correlations Genuine and Spurious in Pearson and Yule. *Statistical Science*. 10. pp. 364–76.

Aldrich, J. (1997) R. A. Fisher and the Making of Maximum Likelihood 1912–22. *Statistical Science*. 12. pp. 162–76.
Aldrich, J. (2010) Mathematics in the London/Royal Statistical Society 1834–1934. *Journal Electronique d'Histoire des Probabilités et de la Statistique*. 6. 1. p. 33.
Aldrich, J. (2012) *A Guide to R. A. Fisher*. [Online] Available from: http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/rafreader.htm. [Accessed 7 Oct 2015.]
Aldrich, J. (2013a) Karl Pearson: A Reader's Guide. [Online] Available from: http://www.economics.soton.ac.uk/staff/aldrich/kpreader.htm. [Accessed 7 Oct 2015.]
Aldrich, J. (2013b) Karl Pearson's *Biometrika*: 1901–36. *Biometrika*. 100. 1. pp. 2–15
Anon (1838) Introduction. *Journal of the Statistical Society of London*. 1. 1.1–5.
Anon (1911) Obituary: Sir Francis Galton, D.C.L., D.Sc., F.R.S. *Journal of the Royal Statistical Society*. 74. 3. pp. 314–20.
Bennett, J. H. (ed.) (1983) *Natural Selection, Heredity, and Eugenics. Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Oxford: Clarendon Press.
Box, J. F. (1978) *R. A. Fisher: The Life of a Scientist*. New York, NY: Wiley.
Bulmer, M. G. (2003) *Francis Galton: Pioneer of Heredity and Biometry*. Baltimore, MD: Johns Hopkins University Press.
Darwin, C. (1868) *The Variation of Animals and Plants under Domestication*. London: Murray.
Elderton, W. P. (1907) *Frequency-Curves and Correlation*. London: Layton.
Elderton, W. P. and Elderton, E. M. (1909) *Primer of Statistics*. London: Black.
Fisher, R. A. (1911) Unpublished paper on "Heredity" (comparing methods of Biometry and Mendelism). Reprinted in Bennett (1983).
Fisher, R. A. (1912) On an Absolute Criterion for Fitting Frequency Curves. *Messenger of Mathematics*. 41. pp. 155–60.
Fisher, R. A. (1915) The Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population. *Biometrika*. 10. 4. pp. 507–21.
Fisher, R. A. (1918) The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*. 52. pp. 399–433.
Fisher, R. A. (1921) On the 'Probable Error' of a Coefficient of Correlation Deduced from a Small Sample. *Metron*. 1. pp. 3–32.
Fisher, R. A. (1922) On the Mathematical Foundations of Theoretical Statistics. *Philosophical Transactions of the Royal Society. A*. 222. pp. 309–68.
Fisher, R. A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.
Fisher, R. A. (1930a) *The Genetical Theory of Natural Selection*. Oxford: Oxford University Press.
Fisher, R. A. (1930b) Inverse Probability. *Proceedings of the Cambridge Philosophical Society*. 26. pp. 528–35.
Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver & Boyd.
Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. Edinburgh: Oliver & Boyd.
Galton, F. (1869) *Hereditary Genius: An Inquiry into its Laws and Consequences*. London: Macmillan.
Galton, F. (1871–2) On Blood Relationship. *Proceedings of the Royal Society of London*. 20. pp. 394–402.
Galton, F. (1875) Prefatory remarks to H. W. Watson, "On the probability of the extinction of families." *Journal of the Anthropological Institute of Great Britain and Ireland*. 4. pp. 138–9.
Galton, F. (1877) Typical Laws of Heredity. *Nature*. 15. pp. 492–5, 512–14, 532–3.
Galton, F. (1883) *Inquiries into Human Faculty and Its Development*. London: Dent.

Galton, F. (1886) Regression towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute*. 15. pp. 246–63.
Galton, F. (1888) Co-relations and their Measurement, Chiefly from Anthropometric Data. *Proceedings of the Royal Society of London*. 45. pp. 135–45.
Galton, F. (1889) *Natural Inheritance*. London: Macmillan.
Galton, F. (1908) *Memories of My Life*. London: Methuen.
Gayon, J. (1998) *Darwinism's Struggle for Survival: Heredity and the Hypothesis of Natural Selection*. Cambridge: Cambridge University Press
Gillham, N. W. (2002) *A Life of Sir Francis Galton: From African Exploration to the Birth of Eugenics*. New York, NY: Oxford University Press.
Hald, A. (1998) *A History of Mathematical Statistics from 1750 to 1930*. New York, NY: Wiley.
Heyde, C. C. and E. Seneta, E. (eds.) (2001) *Statisticians of the Centuries*. New York, NY: Springer.
Hotelling, H. (1930) British Statistics and Statisticians Today. *Journal of the American Statistical Association*. 25. pp. 186–90.
Jeffreys, H. (1939) *Theory of Probability*. Oxford: Oxford University Press.
MacKenzie, D. A. (1981) *Statistics in Britain 1865–1930: The Social Construction of Scientific Knowledge*. Edinburgh: Edinburgh University Press.
Magnello, M. E. (1999) The Non-correlation of Biometrics and Eugenics: Rival Forms of Laboratory Work in Karl Pearson's Career at University College London. (In two parts.) *History of Science*. 37. pp. 79–106, 123–50.
Mazumdar, M. H. (1992) *Eugenics, Human Genetics and Human Failings: The Eugenic Society, its Sources and its Critics in Britain*. London: Routledge.
Neyman, J. (1934) On the Two Different Aspects of the Representative Method (with discussion). *Journal of the Royal Statistical Society*. 97. pp. 558–625.
Neyman, J. and Pearson, E. S. (1928) On the Use of Certain Test Criteria for Purposes of Statistical Inference. Parts I and II. *Biometrika*. 20A. pp. 175–240, 263–94.
Neyman, J. and Pearson, E. S. (1933) On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society*. 231. pp. 298–337.
Pearson, E. S. (1936/8) Karl Pearson: An Appreciation of Some Aspects of his Life and Work. In two parts. *Biometrika*. 28. pp. 193–257, **29**, pp. 161–247.
Pearson, K. (ed.) (1888) *The Common Sense of the Exact Sciences*. London: Kegan Paul, Trench.
Pearson, K. (1892) *The Grammar of Science*. London: Walter Scott.
Pearson, K. (1894) Contributions to the Mathematical Theory of Evolution. *Philosophical Transactions of the Royal Society*. *A*. 185. pp. 71–110.
Pearson, K. (1895) Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material. *Philosophical Transactions of the Royal Society*. *A*. 186. pp. 343–414.
Pearson, K. (1896) Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society of London*. *A*. 187. pp. 253–318.
Pearson, K. (1900) On the Criterion that a Given System of Deviations from the Probable in the Case of Correlated System of Variables is such that it can be Reasonably Supposed to have Arisen from Random Sampling. *Philosophical Magazine*. 50. pp. 157–75.
Pearson, K. (1904) Mathematical Contributions to the Theory of Evolution. XII. On a Generalised Theory of Alternative Inheritance, with Special Reference to Mendel's Laws. *Philosophical Transactions of the Royal Society of London*. *A*. 203. pp. 53–86.

Pearson, K. (1914) *The Life, Letters and Labours of Francis Galton*. Vol. I. Cambridge: Cambridge University Press.
Pearson, K. (1920) The Fundamental Problem of Practical Statistics. *Biometrika*. 13. pp. 1–16.
Pearson, K. (1930) *The Life, Letters and Labours of Francis Galton*. Vol. IIIA. Cambridge: Cambridge University Press.
Pearson, K. (1936) Method of Moments and Method of Maximum Likelihood. *Biometrika*. 28. pp. 34–59.
Pearson, K. and Filon, L. N. G. (1898) Mathematical Contributions to the Theory of Evolution IV. On the Probable Errors of Frequency Constants and on the Influence of Random Selection on Variation and Correlation. *Philosophical Transactions of the Royal Society. A*. 191. pp. 229–311.
Porter, T. M. (2004) *Karl Pearson: The Scientific Life in a Statistical Age*. Princeton, NJ: Princeton University Press.
Provine, W. B. (1971) *The Origins of Theoretical Population Genetics*. Chicago, IL: University of Chicago Press.
Spearman, C. (1904) "General Intelligence," Objectively Determined and Measured. *American Journal of Psychology*. 15. pp. 201–93.
Stigler, S. M. (1986) *The History of Statistics: The Measurement of Uncertainty before 1900*. Cambridge, MA: Harvard University Press.
Stigler, S. M. (2005) Fisher in 1921. *Statistical Science*. 20. pp. 32–49.
Stigler, S. M. (2007) The Epic Story of Maximum Likelihood. *Statistical Science*. 22. 4. pp. 598–620.
Student (1908a) The Probable Error of a Mean. *Biometrika*. 6. pp. 1–25.
Student (1908b) Probable Error of a Correlation Coefficient. *Biometrika*. 6. pp. 302–10.
Wald, A. (1939) Contributions to the Theory of Statistical Estimation and Testing Hypotheses. *Annals of Mathematical Statistics*. 10. pp. 299–326.
Yule, G. U. (1899) An Investigation into the Causes of Changes in Pauperism in England, Chiefly during the Last Two Intercensal Decades. *Journal of the Royal Statistical Society*. 62. 2. pp. 249–95.
Yule, G. U. (1902) Mendel's Laws and their Probable Relations to Intra-racial Heredity. *New Phytologist*. 1. pp. 193–207, 222–38.
Yule, G. U. (1911) *Introduction to the Theory of Statistics*. London: Griffin.
Yule, G. U. (1936) Karl Pearson. In *Obituary Notices of Fellows of the Royal Society of London*. 2. 1. pp. 74–110.

CHAPTER 7

THE ORIGINS OF PROBABILISTIC EPISTEMOLOGY

Some Leading 20th-century Philosophers of Probability

MARIA CARLA GALAVOTTI

The notion of probability received great attention from 20th-century mathematicians and philosophers alike. This chapter focuses on a number of thinkers who not only devoted great efforts to the notion of probability and its foundations, but also developed a thoroughly probabilistic epistemological perspective. Special attention will be paid to Hans Reichenbach, Harold Jeffreys, and Bruno de Finetti. Although these authors embraced diverging interpretations of probability, namely frequentism in the case of Reichenbach, logicism in the case of Jeffreys, and subjectivism in the case of de Finetti, they shared the conviction that probability is an essential ingredient not just of science, but of human knowledge at large, and laid the foundations of a probabilistic approach to epistemology that today is mainstream.

7.1 Some Background

To start with, it is not out of place to introduce the main interpretations of probability that were the focus of 20th-century debate[1]. In the course of the 19th century, the "classical" view of probability developed by Pierre Simon de Laplace (1749–1827) was gradually superseded by two major trends: logicism and frequentism. *Logicism* is the thesis that the theory of probability belongs to logic, and probability is a logical relation between two propositions, of which one states a given hypothesis and the other describes the evidence supporting it. The frequency interpretation defines probability as the limit of the relative frequency of a given attribute observed in the initial part of an indefinitely long sequence of repeatable events. Logicism was embraced by the Czech Bernard Bolzano (1781–1848) and by a number of

[1] See Galavotti (2005) for an extensive introduction to the debate on the various interpretations of probability and their beginnings, and to the work of the authors mentioned in this chapter.

British authors including Augustus De Morgan (1806–71), George Boole (1815–64), and William Stanley Jevons (1835–82).

The fortunes of logicism in the 20th century depend at least in part on the Cambridge logician William Ernest Johnson (1858–1931), whose lectures at King's strongly influenced a number of intellectuals including Frank Ramsey, John Maynard Keynes, and Harold Jeffreys. Johnson has the merit of having introduced the property known as *exchangeability*, the subject of his "Permutation postulate". Johnson's discovery of this result left some trace in Ramsey's work, but remained almost ignored until the statistician Irving John Good called attention to it in his monograph *The Estimation of Probabilities. An Essay on Modern Bayesian Methods*, which opens with the following words: "This monograph is dedicated to William Ernest Johnson, the teacher of John Maynard Keynes and Harold Jeffreys" (Good 1965, p. v). Good made extensive use of Johnson's ideas in a Bayesian framework[2].

An important step in the development of logicism was the publication in 1921 of John Maynard Keynes' *Treatise on Probability*. This book is a systematic study of probability, arguing for a moderate form of logicism which assigns an important role to intuition and individual judgment. Keynes also ascribes an important role to *analogy*, and calls attention to the importance of the notion of *weight*. Like probability, the weight of inductive arguments varies according to the amount of evidence, but whereas probability is affected by the proportion between favourable and unfavourable evidence, weight increases as relevant evidence – taken as the sum of positive and negative observations – increases. This notion was subsequently extensively investigated by authors of Bayesian orientation.

Logicism reached its apex with Rudolf Carnap (1891–1970). Starting from the admission that there are two concepts of probability: namely *probability*$_1$, or *degree of confirmation*, and *probability*$_2$, or probability as *frequency*, Carnap set himself the task of developing probability$_1$ as the object of *inductive logic* in his monumental *Logical Foundations of Probability*, followed by a number of publications, some of which appeared posthumously. Inductive logic is constructed as an axiomatic system, formalized within a first-order predicate calculus with identity, which applies to measures of confirmation defined on the semantic content of statements. Carnap regards inductive logic as a rational basis for decisions, since it allows for best estimates based on the given evidence. Logical probability is analytical and objective: in the light of the same body of evidence there is only one rational (correct) probability assignment. Carnap devised a *continuum of inductive methods* characterized as a blend of two components: one purely logical and one purely empirical. A privileged position in the continuum is assigned to the functions that Carnap called *symmetric*, having the property of exchangeability[3]. The indisputable merit of Carnap's approach is that it clarified the conceptual presuppositions of probabilistic inferences which become fully explicit once they are embodied in axioms. This positive aspect of rational reconstruction performed by logical tools is counterbalanced by the shortcoming of its awkward formalism, which has not appealed to statisticians and scientists, with the drawback that Carnap's work on probability has barely gone beyond the restricted circle of philosophers of science.

Since Carnap's methods belong to the broader family of Bayesian methods, discussion of these tools has to a certain extent mingled with that of Bayesian confirmation, which

[2] See Zabell (1982) for a historical and theoretical discussion of Johnson's formulation of the notion of exchangeability and its relevance for Bayesian statistics.

[3] See Carnap (1952).

represents the mainstream of the literature on probabilistic confirmation. This was the direction taken by Richard Jeffrey. After having studied under Carnap in Chicago in the early 1950s and followed his entire production[4], Jeffrey subsequently abandoned Carnap's "rationalist Bayesianism" to turn to Bruno de Finetti's subjectivism.

Other representatives of 20th-century logicism include Ludwig Wittgenstein (1889–1951) and Friedrich Waismann (1896–1959), whose ideas anticipated Carnap's inductive logic. A slightly different version of logicism was propounded by Harold Jeffreys, whose work is discussed in Section 7.3.

Frequentism was developed in the 19th century by Robert Leslie Ellis (1817–59) and John Venn (1834–1923), to reach its climax with the rigorous formulation of Richard von Mises (1883–1953) who clarified its presuppositions and fixed the boundaries of its range of application. Von Mises' theory is grounded on the notion of *collective*, denoting an indefinitely long sequence of observations of a mass phenomenon exhibiting frequencies tending to a limit. The distinctive feature of a collective is *randomness*, defined through the operative notion of *place selection*. This points to a method of extracting sub-sequences from the original sequence by taking into account only the places occupied by members in the sequence, not the attributes characterizing them. Each place selection is defined by a rule that states for any element of the sequence whether it ought to be made part of the sub-sequence or not. For example, the sub-sequence obtained by choosing all members whose place number in the sequence is a prime number would satisfy von Mises' criterion. Randomness is identified with "insensitivity to place selection". It obtains when the limiting values of the relative frequencies in a given collective are not affected by any of all the possible selections that can be made. In addition, the limiting values of the relative frequencies, in the sub-sequences obtained by place selection, equal those of the original sequence. By defining randomness in terms of insensitivity to *all* possible place selections, von Mises embraces an absolute, unrestricted notion of randomness, which raised various objections[5].

The theory of probability is re-stated by von Mises in terms of collectives, through the operations of "selection", "mixing", "partition", and "combination". By means of this conceptual machinery, von Mises sought to endow probability with a solid foundation, both empirical and objective. From von Mises' perspective probability can be applied to scientific phenomena only insofar as they can be reduced to "chance mechanisms" having the features of collectives. Consequently, probability evaluations referred to single occurrences of phenomena are not allowed. This raises the difficulty known as the "problem of the single case".

A further problem of applicability arises in connection with the use of infinite sequences, which collides with von Mises' intention to develop an empirical notion of probability, operationally reducible to a measurable quantity. While holding that probability as an idealized limit can be compared to other limiting notions encountered in science, such as velocity or density, von Mises admits that a problem of applicability may arise in connection with the relationship between the sequences of observations, which are obviously finite, and the infinite sequences postulated by the theory. Such a relationship involves a twofold inductive passage: from observation to theory and vice versa. Von Mises does not go into the details of this passage, claiming that it "is not logically definable, but is nevertheless

[4] Jeffrey edited Carnap's last manuscripts and collected them in Carnap and Jeffrey (eds.) (1971) and Jeffrey (ed.) (1980).

[5] See Martin-Löf (1969) for a survey of the problems raised by von Mises' theory of randomness.

sufficiently exact in practice" (von Mises 1928, English edition 1939/1957, p. 85). The impossibility of formulating single-case probability assignments and the lack of a fully developed theory of induction stimulated Hans Reichenbach to work out an alternative version of frequentism, to be recollected in Section 7.2.

In spite of the problems it leaves open, frequentism received such a formidable impulse from the work of von Mises that it became the "official" interpretation of probability accepted by scientists operating in all areas of the natural sciences. It is noteworthy that von Mises forcefully affirms the probabilistic character of knowledge and endorses indeterminism. In addition, he denies the principle of causality, claiming that within modern science recourse to statistical methods has gradually superseded causal talk.

Another interpretation of probability took shape in the first decades of the 20th century, namely *subjectivism*. According to subjectivists probability is the *degree of belief* entertained by a person in a state of uncertainty regarding the occurrence of an event on the basis of the information available. Degree of belief is assumed as a primitive notion, which has to be given an operative definition specifying how it can be measured. A well-known way to achieve this goal – though by no means the only one – is the method of bets, with a long-standing tradition dating back to the 17th century.

Anticipated by the British astronomer William Donkin (1814–69) and the French mathematician Émile Borel (1871–1956), the subjective approach was given a sound basis by Frank Plumpton Ramsey (1903–30), whose most important contribution to the subject is "Truth and Probability", read at the Moral Sciences Club in Cambridge in 1926 and published posthumously in 1931 as part of the collection *The Foundations of Mathematics and Other Logical Essays* edited by Richard Bevan Braithwaite[6]. Ramsey adopts a definition of degree of belief based on preferences determined on the basis of an individual's expectation of obtaining certain goods, not necessarily of a monetary kind, and specifies a set of axioms fixing a criterion of *coherence*. According to betting terminology, coherence ensures that, if used as the basis of betting ratios, degrees of belief should not lead to a sure loss. Ramsey states that coherent degrees of belief satisfy the laws of probability. Thereby coherence became the cornerstone of the subjective interpretation of probability. As Ramsey put it, within this perspective the laws of probability "do not depend for their meaning on any degree of belief in a proposition being uniquely determined as the rational one; they merely distinguish those sets of beliefs which obey them as consistent ones" (Ramsey 1990, p. 78). Coherence is the only condition that degrees of belief should obey; insofar as a set of degrees of belief is coherent, there is no further demand of rationality to be met. This marks a sharp difference between subjectivism and the logical interpretation of probability.

Contrary to the widespread opinion that Ramsey was a dualist with regard to probability, in the last years of his life he put forward a view of *chance* and *probability in physics* fully compatible with his subjective interpretation of probability as degree of belief. Without going into the details of his view of the matter, he defines chance as degree of belief of a special kind, whose peculiarity is that of being always referred to a "system of beliefs", rather than to the beliefs of certain agents. A distinctive feature of the systems to which chance is referred is that it includes laws and other statements describing the behaviour of phenomena under consideration, such as correlation statements. Such laws, in conjunction with the

[6] Reprinted in Ramsey (1990); quotations in this paper are from that edition.

empirical knowledge possessed by the system's users, entail degrees of belief representing "chances", to which the actual degrees of belief held by users should approximate. This notion applies to "chance phenomena" such as games of chance, whereas "probability in physics" is defined as chance referred to a more complex system, including reference to scientific theories. In other words, probabilities occurring in physics are derived from physical theories, and their objective character descends from the objectivity peculiarly ascribed to theories that are commonly accepted as true[7]. Ramsey's idea that within the framework of subjective probability one can make sense of an "objective" notion of physical probability has passed almost unnoticed. It is, instead, an important contribution to the subjective interpretation and its application to science.

A further interpretation of probability put forward in the 20th century is the so-called *propensity theory*, proposed in the 1950s by Karl Raimund Popper (1902–94) to solve the problem of single-case attributions arising within quantum mechanics. Probability as propensity is a *property of the experimental arrangement* suited to reproduction over and over again to form a sequence[8]. Starting from the early 1980s Popper came to see propensities as "weighted possibilities", or expressions of the tendency of a given experimental set-up to realize itself upon repetition, emphasizing single experimental arrangements rather than sequences of generating conditions. In so doing, he laid down the so-called "single-case propensity interpretation". Of crucial importance in this connection is the distinction between "probability statements" expressing propensities, which are statements about frequencies in virtual sequences of experiments, and "statistical statements" expressing relative frequencies observed in actual sequences of experiments, which are used to test probability statements. Popper regards propensities as physically real and metaphysical (they are non-observable properties), and they form the basis of his indeterministic view of reality[9]. After Popper's work the propensity theory of probability has received some consensus among philosophers of science[10].

7.2 Hans Reichenbach's All-encompassing Inductivism

A member of the "Berlin Society for empirical philosophy", Hans Reichenbach (1891–1953) developed a version of frequentism that strays in many respects from that of von Mises, and based on it a deeply probabilistic epistemology rooted in the conviction that probability, not truth, can substantiate a viable theory of knowledge and a reconstruction of science in tune with scientific practice. In an address delivered to the *Neuvième Congrès International de Philosophie* (Paris, 1937) Reichenbach maintained that "the ideal of an absolute truth is an unrealizable phantom; certainty is a privilege pertaining only to tautologies, namely those propositions which do not convey any knowledge" (Reichenbach 1937, p. 90. My

[7] For more on Ramsey's notion of chance and probability in physics see Galavotti (1995) and (1999).
[8] See Popper (1959).
[9] See Popper (1990).
[10] See Gillies (2000) for more on the propensity theory.

translation). When dealing with matters of experience, truth can at best represent the limiting case of probability, namely "a special case in which the probability value is near to one or zero. It would be illusory to imagine that the terms 'true' or 'false' ever expressed anything other than high or low probability values" (Reichenbach 1936, p. 156). The pivotal role played by probability within science and knowledge in general descends from the predictive character of probabilistic statements. Since claims about the future are uncertain, the theory of knowledge requires the theory of probability.

In this spirit, Reichenbach rejected both the verifiability criterion of meaning characterizing the early days of logical empiricism, and the theory of partial definability subsequently put forward by Carnap[11]. By way of contrast, he heralded a probabilistic theory of meaning which "abandoned the program of defining 'the meaning' of a sentence. Instead, it merely laid down two principles of meaning; the first stating the conditions under which a sentence has meaning; the second the conditions under which two sentences have the same meaning" (Reichenbach 1951, p. 47). Based as it is on a probabilistic theory of meaning, Reichenbach's theory of knowledge is probabilistic all the way through, and it rests on the frequency interpretation of probability.

A major feature of Reichenbach's perspective is his concern for scientific practice. This imbues his version of frequentism, which strays from that of von Mises because it allows for single-case probability attributions, admits a weaker concept of randomness, and includes a theory of induction and an argument for its justification. With an eye to practical applications, Reichenbach adopted a notion of randomness not based on an unrestricted invariance domain, but relative to a limited domain of selections "not defined by mathematical rules, but by reference to physical (or psychological) occurrences" (Reichenbach 1935, English edition 1949/1971, p. 150). In the same spirit, he weakened von Mises' definition of limit, by introducing the notion of *practical limit* "for sequences that, in dimensions accessible to human observation, converge sufficiently and remain within the interval of convergence", adding that "it is with sequences having a practical limit that all actual statistics are concerned" (*ibid*., pp. 347–8).

Reichenbach regarded probability as inextricably intertwined with induction, and his formulation of the frequentist canon for assessing probabilities reflects that conviction. He calls the method by which probabilities are obtained *induction by enumeration*. This consists in calculating the relative frequency of a certain attribute in an initial section of a sequence which is being considered, and *assuming* that the observed frequency will hold approximately, or within certain limits of exactness, for any prolongation of the sequence. Reichenbach argues that such an assumption "formulates the *principle of induction*" (Reichenbach 1938, p. 340), which he regards as inseparable from the theory of probability. It is precisely the unification of induction and probability that suggests the procedure which he calls *determination a posteriori* of probability. This is stated in more precise terms by the *Rule of induction*, which says that "if the sequence has a limit of the frequency, there must exist an n such that from there on the frequency f^i ($i > n$) will remain within the interval $f^n \pm \delta$, where δ is a quantity that we can choose as small as we like, but that, once chosen, is kept constant. Now if we posit that the frequency f^i will remain within the interval $f^n \pm \delta$, and we correct this posit for greater n by the same rule, we must

[11] See Carnap (1936–7).

finally come to the correct result" (Reichenbach 1935, English edition 1949/1971, p. 445). Reichenbach's formulation makes it explicit that the definition of probability as limiting frequency requires an inductive step, and that every probability attribution is a wager, or what he calls a *posit*.

The notion of posit plays a crucial role within Reichenbach's epistemology, where it serves to bridge the gap between the probability of sequences and single-case attributions. A posit regarding a single occurrence of an event receives a *weight* from the probabilities attached to the reference class to which the event in question has been assigned, which must obey a criterion of *homogeneity*, namely, it should be chosen in such a way as to include as many cases as possible similar to the one under consideration, excluding dissimilar ones. Homogeneity is obtained through successive partitions of the reference class by means of statistically relevant properties. A reference class is homogeneous when it cannot be further partitioned in this way. Reichenbach recommends choosing "the narrowest class for which we have reliable statistics" (Reichenbach 1938, p. 316).

Posits differ depending on whether they are made in a situation of primitive or advanced knowledge. Reichenbach calls the state of knowledge in which some knowledge of probabilities is available *advanced*, and that in which such knowledge is unavailable *primitive*. Within primitive knowledge, use of the rule of induction yields probability values, while in the context of advanced knowledge use is made of the calculus of probabilities. Scientific hypotheses are confirmed within the framework of advanced knowledge by means of Bayes' rule. The pivotal role assigned in this connection to the Bayesian method is a distinctive feature of Reichenbach's approach, according to which prior probabilities are determined on the basis of frequencies. Accordingly, Reichenbach qualifies as an "objective Bayesian".

All posits are characterized by what Reichenbach calls a *weight*, but while *appraised* posits, made in the context of advanced knowledge, have a definite weight, *blind* posits, made in the context of primitive knowledge, have unknown weight and are approximate in character. However, if the sequence has a limit blind posits can be corrected. Reichenbach calls the procedure that starts with blind posits and goes on to formulate appraised posits that become part of a complex system the *method of concatenated inductions*. By virtue of the convergence assured by the Rule of induction this method qualifies as self-correcting. For this reason, Reichenbach regards it as the essence of science: "the system of science [...] must be regarded as a system of posits" (Reichenbach 1935, English edition 1949/1971, p. 469). As echoed by the title of his book *Experience and Prediction*, for Reichenbach the core of scientific method is the interplay between these two conceptual operations, and the method of concatenated inductions bridges them by allowing the shift from experience of frequencies to predictions of probabilities.

The approximate character of blind posits offers the ground for Reichenbach's argument for the justification of induction. This has a pragmatic flavour, being based ultimately on the notion of success, and consists in the claim that "the rule of induction is justified as an instrument of positing because it is a method of which we know that if it is possible to make statements about the future we shall find them by means of this method" (*ibid.*, p. 475). In other words, we know that by making and correcting posits we will eventually reach success, *if* the considered sequence has a limit. As Salmon observed, the strength of Reichenbach's argument is that it does not require the presupposition that there exists an order in nature. In fact, if nature is uniform other methods of inference may work, but this cannot be shown in

the same way as it is for the Rule of induction; if nature is not uniform the Rule of induction will not work, but other methods will not work either, for if some other method led us to make correct predictions habitually, this fact alone would constitute a uniformity to which induction could be applied. In Salmon's words: "it should be clear that a solution such as Reichenbach's to the problem of justification of induction avoids the necessity of any such *assumption* as the principle of uniformity of nature. The whole force of the justification is that the use of induction is reasonable whether or not nature is uniform, whatever may be meant by the assertion 'Nature is uniform'" (Salmon 1953, p. 48). At the same time, Salmon regards the fact that Reichenbach's argument justifies a whole class of asymptotic rules as a limitation, and made an attempt to restrict it to the Rule of induction alone[12].

It is noteworthy that Reichenbach attaches a probabilistic meaning to the notion of causality, pioneering a major trend of research in philosophy of science. The basic intuition underpinning his approach is to ground *causal relevance* on *statistical relevance*, subject to restrictions devised to entertain the fact that causal relations are asymmetrical, whereas statistical relevance relations are symmetrical. Reichenbach grounds the asymmetry of time on the asymmetry of causal relations, and develops a causal theory of time in which the direction of time, as well as the notion of temporal priority, is defined on the basis of causal asymmetry and antecedence[13]. The fundamental ingredient in this connection is the *Principle of the common cause*. In brief, this principle states that if two events of a certain kind happen jointly – though in different places – more often than would be expected were they independent, this apparent coincidence should be explained in terms of a common causal antecedent. This principle was resumed by Wesley Salmon to underpin a theory of scientific explanation inspired by Reichenbach's idea that "the causal structure of the universe can be comprehended with the help of the concept of *probable determination* alone" (Reichenbach 1925, English edition 1978, vol. 2, p. 83)[14].

Reichenbach judged the traditional distinction between probabilistic regularities and causal laws as "only superficial" because "probability laws and causal laws are logical variations of one and the same type of regularity" (Reichenbach 1930, English edition 1959/1978, vol. 2, p. 335). General statements about natural phenomena are deemed subject to various conditions, and involve some degree of uncertainty. As Reichenbach put it: "the characterization of the causal laws of nature as strict laws is justified only for certain schematizations. When all causal factors are known, then an effect can be predicted with certainty; such an idealization would be irrelevant for science, without the addition of further assumptions. [...] Actually, we can only maintain that it is highly probable that future events will lie within certain limits of exactness" (*ibid.*).

Reichenbach's epistemology represents a unique blend of empiricism and pragmatism, originating from an unshakable trust in empiricism combined with the deeply felt conviction that our knowledge is uncertain and relies on induction. He sets himself the goal of building an epistemological perspective which, although probabilistic, has a basis sound enough to guarantee objectivity, and identifies this solid bedrock with the frequency interpretation of probability. Reichenbach is altogether concerned that the

[12] For more on Reichenbach's justification of induction and Salmon's attempt to refine it see Galavotti (2011a) also containing a detailed discussion of Reichenbach's inductivism.

[13] See Reichenbach (1956) for the author's causal theory of time.

[14] See Salmon (1971), (1984) and (1998) for further reading.

rational reconstruction of science should never lose sight of scientific practice nor of everyday life. It is somewhat ironical that in spite of his constant concern for practice, his perspective fails precisely on application grounds. Severe problems arise in connection with identifying the homogeneous reference class to which single-case probability attributions should be referred. Further problems beset the criterion of convergence on which the Rule of induction is based, and indeed the whole method of concatenated inductions at the core of Reichenbach's epistemology. The problem is that with respect to any given sequence one cannot predict how far one must go before obtaining evaluations – posits in Reichenbach's terminology – that fulfil a desired degree of accuracy[15]. Reichenbach's version of frequentism has not appealed much to statisticians and scientists, who have for the most part turned to von Mises' theory of probability. On the other hand, Reichenbach's probabilistic epistemology is very insightful and suggestive of new developments, as testified, among other things, by Salmon's seminal work on explanation.

7.3 Harold Jeffreys' Probabilistic Epistemology

The British geophysicist Harold Jeffreys (1891–1989), a pioneer of the study of the Earth, gave outstanding contributions to probability and statistics and developed an original epistemological perspective[16]. Working in a field like geophysics, where massive data were not available, Jeffreys was suspicious of the frequency theory, at the time the most accredited approach to probability in physics and science at large. By contrast, he embraced the logical interpretation of probability that was prominent in Cambridge thanks to William Ernest Johnson and John Maynard Keynes. Although sharing some of Keynes' traits and also Carnap's logicism, Jeffreys strays from these authors in important respects, resembling the subjectivism of Ramsey and de Finetti. Jeffreys' epistemology is developed in a number of articles in addition to the two volumes *Scientific Inference* (1931) and *Theory of Probability* (1939), both including extensive discussions of philosophical matters. At its roots lies the conviction that probability is "the most fundamental and general guiding principle of the whole of science" (Jeffreys 1931, p. 7).

Jeffreys takes probability as "a purely epistemological notion" (Jeffreys 1955, p. 283) expressing the degree of belief entertained in the occurrence of an event on the basis of a given body of evidence, and holds that the theory of probability "should start from definite postulates and the consequences should follow from the postulates" (Jeffreys 1963, p. 409). Probability is grounded on a principle, stated by way of an axiom, which says that probabilities are comparable: "given p, q is either more, equally, or less probable than r, and no two of these alternatives can be true" (Jeffreys 1939/1961, p. 16). He then demonstrates that the fundamental properties of probability functions follow from this assumption. By so doing, Jeffreys qualified as one of the first to establish the rules of

[15] See Eberhardt and Glymour (2011) for a discussion of this problem.

[16] See the memoir in Cook (1990) on Jeffreys' life and scientific achievements. See also Galavotti (2003a), from which this section is partly taken.

probability, namely convexity, addition and multiplication, from basic presuppositions. A similar accomplishment was attained – totally independently – by Ramsey and de Finetti.

A central feature of Jeffreys' viewpoint is the tenet that *reasonable* degrees of belief are uniquely determined by evidence, so that given a set of data "a proposition q has in relation to these data one and only one probability. If any person assigns a different probability, he is simply wrong" (Jeffreys 1931, p. 10). There follows that a satisfactory theory has to account for the uniqueness of probability values: "our main postulates are the existence of unique reasonable degrees of belief, which can be put in a definite order" (Jeffreys 1939, p. 36). This conviction puts Jeffreys with the logicists and marks a crucial divergence from subjectivism, a divergence which is described by de Finetti as that "between necessarists who affirm and subjectivists who deny that on logical grounds it is possible to specify *one* evaluation of probability, which is seen as *objectively* privileged and 'correct'" (de Finetti 1970, English edition 1975, p. 391)[17].

A crucial role within Jeffreys' epistemology is assigned to Bayes' rule, of which he says that it "is to the theory of probability what Pythagoras' theorem is to geometry" (*ibid.*, p. 19). Jeffreys was led to embrace the Bayesian method by his own work as a practising scientist in which he faced problems of inverse probability, having to explain experimental data by means of different hypotheses, or to evaluate general hypotheses in the light of changing data. Jeffreys started working on these problems around 1920 together with Dorothy Wrinch, a mathematician who later made important contributions to the study of protein structure, drawing the lines of an inductivist programme that he kept firm throughout his life[18]. This is grounded on the assumption that all quantitative laws form an enumerable set, and their probabilities form a convergent series. This assumption allows for the assignment of significant prior probabilities to general hypotheses. In addition, Jeffreys and Wrinch formulated a *simplicity postulate*, according to which simpler laws are assigned a greater prior probability[19]. According to its proponents, this principle corresponds to the practice of testing possible laws in decreasing order of simplicity.

Jeffreys' views on probability strayed from those upheld by the majority of statisticians of his time, which led him to disagree with Ronald Fisher[20]. As Dennis Lindley observed, "Jeffreys considered probability as the appropriate description for all uncertainty, whereas statisticians usually restrict its use to the uncertainty associated with data. The collection of data can ordinarily be repeated and the frequency concept of probability is relevant. Scientific laws admit of no repetition and their uncertainty resides in the belief that they might be true" (Lindley 1991, p. 11). This attitude is reflected by Jeffreys' Bayesianism, which derives an objective flavour from his "necessarist" attitude. While diverging from other kinds of Bayesianism, like de Finetti's, this aspect of Jeffreys' perspective anticipates the recent trend of research called "objective Bayesianism", which also draws inspiration from the work of Richard Cox and Edwin Jaynes[21].

The need to assume a notion of probability suitable for scientific applications leads Jeffreys to criticize the subjective interpretation of probability put forward by Ramsey, with whom he

[17] The claim that Jeffreys was a "necessarist" is rebutted in Zellner (1982).

[18] See Howie (2002) for a detailed reconstruction of the genesis of Jeffreys' Bayesianism and his cooperation with Dorothy Wrinch.

[19] A discussion of the simplicity postulate is to be found in Howson (1988).

[20] See Jeffreys (1933) and (1934) for Jeffreys' arguments against Fisher.

[21] For more on this see Williamson (2009).

shared various interests but never discussed probability[22]. Jeffreys deems Ramsey's a "theory of expectation" good "for business men". This is not meant as an expression of contempt, for "we have habitually to decide on the best course of action in given circumstances, in other words to compare the expectations of the benefits that may arise from different actions; hence a theory of expectation is possibly more needed than one of pure probability" (Jeffreys 1939/1961, p. 326); still, science requires a notion of "pure probability", not a subjective notion in terms of preferences based on expectations.

A point of agreement between Jeffreys and subjectivists was the idea of "unknown probability", rejected by both. Jeffreys' opposition to this idea, which was instead retained by John Maynard Keynes, is reaffirmed in the second edition of *Theory of Probability*, where the author claims that "unknown probability" involves either a confusion between the statements "I do not know x" and "I do not know the probability of x", or the identification of prior probability and frequency, but "no coherent theory can be made until we are rid of both" (*ibid.*, pp. 406–7). The above argument testifies to Jeffreys' aversion to frequentism, to which he opposes arguments akin to those adopted by subjectivists, essentially based on the conviction that "any definition of probability that attempts to define probability in terms of infinite sets of possible observations" has to be rejected, "for we cannot in practice make an infinite number of observations" (*ibid.*, p. 11). From these premises he draws the conclusion that "no 'objective' definition of probability in terms of actual or possible observations, or possible properties of the world, is admissible" (*ibid.*). Actually, this relationship is reversed within Jeffreys' perspective, where probability comes before the notions of objectivity, reality and external world.

Jeffreys' probabilistic epistemology features a constructivist attitude according to which the notions of *objectivity* and *reality* are established by inference from experience by means of statistical methodology, which is the essence of scientific method. In the "Addenda" to the 1937 edition of *Scientific Inference* we read that "the introduction of the word 'objective' at the outset seems [...] a fundamental confusion. The whole problem of scientific method is to find out *what* is objective (I prefer the word *real*) and all we can do is to examine possibilities in turn and see how far they coordinate with observations. Without a theory of knowledge this question has no answer; and with it the other question 'what is the external world really like?' can be left to metaphysics, where it belongs" (Jeffreys 1931/1937, p. 255). A similar assertion is contained in *Theory of Probability*, where Jeffreys writes: "I should query whether any meaning can be attached to 'objective' without a previous analysis of the process of *finding out* what is objective" (Jeffreys 1939, p. 336). The process in question is inductive and probabilistic, it originates in our experience and proceeds step by step to the construction of abstract notions lying beyond phenomena. Although such notions cannot be described in terms of observables, they are nonetheless admissible and useful, because they permit "co-ordination of a large number of sensations that cannot be achieved so compactly in any other way" (Jeffreys 1931/1973, p. 190). Empirical laws, commonly taken as "objective statements", are established through an inductive step, for it is only after the rules of induction "have compared it with experience and attached a high probability to it as a result of that comparison" that a general proposition can become a law. In this procedure lies "the only scientifically useful meaning of 'objectivity'" (Jeffreys 1939, p. 336).

[22] According to Howie and Lindley, Jeffreys found out about Ramsey's work on probability only after Ramsey's death in 1930; see Howie (2002), p. 117 and Lindley (1991), p. 13.

The same holds for the strictly related notion of reality; it is once again the theory of probability, taken to include statistical methodology, that allows for the only concept of "reality" that proves useful to the establishment of scientific knowledge. Such a notion obtains when some scientific hypotheses receive from the data a probability that is so high, that on their basis one can draw inferences, whose probabilities are practically the same as if the hypotheses in question were certain. Hypotheses of this kind are taken as certain in the sense that all their parameters "acquire a permanent status". At that point the associations expressed by the hypotheses in question can be asserted "as an approximate rule". To be sure, the "scientific notion of reality" so obtained "is not an *a priori* one at all, but a rule of method that becomes convenient whenever we can replace an inductive inference by an approximate deductive one. The possibility of doing this in any particular case is based on experience" (Jeffreys 1937, p. 69).

Jeffreys retains an equally empirical and constructivist view of *causality*. His proposal is to substitute the general formulation of the "principle of causality" with "causal analysis", as performed by means of statistical methodology. This starts by considering all the variations observed in a given phenomenon at random, and proceeds to detect correlations which allow for predictions and descriptions whose precision is directly correlated to their agreement with observations. This procedure leads to asserting laws, which are eventually accepted because "the agreement (with observations) is too good to be accidental" (*ibid.*, p. 62). Within scientific practice, the principle of causality is "inverted": "instead of saying that every event has a cause, we recognize that observations vary and regard scientific method as a procedure for analysing the variation" (Jeffreys 1931/1957, p. 78). The deterministic version of the principle of causality is thereby discarded, for "it expresses a wish for exactness, which is always frustrated, and nothing more" (Jeffreys 1937, pp. 63–4). Regarding general propositions, laws and causality, Jeffreys' approach shares the pragmatic flavour characterizing Ramsey's views of these matters, the main difference being that Ramsey's approach is more strictly analytic, whereas Jeffreys grounds his arguments on probabilistic inference and statistical methodology alone.

Furthermore, Jeffreys and Ramsey share the conviction that within an epistemic interpretation of probability there is room for notions like *chance* and *physical probability*. Jeffreys regards the notion of chance as the "limiting case" of everyday probability assignments. Chance occurs in those situations in which "given certain parameters, the probability of an event is the same at every trial, no matter what may have happened at previous trials" (Jeffreys 1931/1957, p. 46). He also considers the possibility of extending the realm of epistemic probability to a robust notion of "physical probability" of the kind encountered in quantum mechanics, calling attention to those fields where "some scientific laws may contain an element of probability that is intrinsic to the system" (Jeffreys 1955, p. 284). Unlike the probability (chance) that a fair coin falls heads, intrinsic probabilities do not belong to our description of phenomena, but to the theory itself. Jeffreys claims to be "inclined to think that there may be such a thing as intrinsic probability." Whether there is or not – he adds – "it can be discussed in the language of epistemological probability" (*ibid.*, p. 284). As we have seen, a similar attitude is taken by Ramsey, who is likely to have been influenced by Jeffreys in this connection.

Jeffreys' constructivist and pragmatist epistemology carries the conviction that science is fallible and that there is a continuity between science and everyday life, since inductive inference obeys rules of consistency "that will agree with ordinary behaviour" (Jeffreys

1936, p. 331). He goes as far as admitting that empirical information can be "vague and half-forgotten" (Jeffreys 1931/1973, p. 406), anticipating the literature on probability kinematics, which contemplates the application of Bayesian inference to uncertain information[23].

7.4 Bruno de Finetti's Radical Probabilism

Together with Ramsey, Bruno de Finetti (1905–85) is the reputed founder of the subjective interpretation of probability. To the definition of probability as degree of belief subject to the only constraint of coherence, de Finetti added the notion of *exchangeability*, which can be regarded as the decisive stepping-stone towards modern subjectivism. Combined with Bayes' rule, exchangeability gives rise to the inferential methodology which is at the root of so-called neo-Bayesianism. In addition to seminal contributions to probability, statistics, and other fields such as economic theory, de Finetti developed a probabilistic epistemology resulting from an original combination of pragmatism and the kind of empiricism that is today called "anti-realism"[24]. Richard Jeffrey labelled de Finetti's viewpoint "Radical probabilism"[25], whereas the author called his own position "subjective Bayesianism" to stress the fundamental role assigned to Bayes' rule. It revolves around a conception of scientific knowledge as a product of human activity ruled by probability rather than truth.

De Finetti's article "Probabilismo", which he regarded as his philosophical manifesto, traces his own philosophy back to Mach's phenomenalism, a view which also influenced Harold Jeffreys[26]. Also influential on de Finetti were the Italian pragmatists Giovanni Vailati, Antonio Aliotta and Mario Calderoni. "Probabilismo" contains a refusal of the notion of truth, and the related view that there are "immutable and necessary" laws. As de Finetti put it: "no science will permit us to say: this fact will come about, it will be thus and so because it follows from a certain law, and that law is an absolute truth. Still less will it lead us to conclude sceptically: the absolute truth does not exist, and so this fact might or might not come about, it may go like this or in a totally different way, I know nothing about it. What we can say is this: *I foresee* that such a fact will come about, and that it will happen in such and such a way, because past experience and its scientific elaboration by human thought make this forecast seem reasonable to me" (de Finetti 1931, English edition 1989, p. 170). As echoed by the title of one of his major works: "La prévision: ses lois logiques, ses sources subjectives" (1937), the primary goal of science is to make good forecasts. Use of probability is what makes forecasting possible, and since a forecast is but the product of a person's experience and convictions, together with the information available to him, the proper tool for prediction is the subjective theory of probability.

[23] Jeffrey (1992a) and (2004) are landmarks in the vast literature on the topic. For some recent contributions see Bovens and Hartmann, eds. (2007).

[24] For a sketchy intellectual autobiography of de Finetti's the reader is addressed to de Finetti (1982). For a comparison between Ramsey's and de Finetti's versions of subjectivism see Galavotti (1991).

[25] See "Introduction: Radical Probabilism", in Jeffrey (1992a), pp. 1–13, and Jeffrey (1992b).

[26] See Jeffreys (1931), p. 225.

De Finetti regards probability as degree of belief "as actually held by someone, on the ground of his whole knowledge, experience, information" (de Finetti 1968, p. 45). Probability is a primitive notion that has to be endowed with an operational definition in order to be measured. The betting scheme is a classical means of devising an operational definition of probability: the degree of probability assigned by an individual to a certain event is identified with the betting quotient at which he would be ready to bet a certain sum on its occurrence, and probability is defined as the fair betting quotient he would attach to his bets. Like Ramsey, de Finetti regards coherence as the fundamental and unique criterion of rationality, and spells out an argument to the effect that coherence is a sufficient condition for the fairness of a betting system, showing that a coherent gambling behaviour satisfies the principles of probability calculus, which can be derived from the notion of coherence itself. It should be stressed that the betting scheme is not the only way to provide an operative definition of probability, which can also be produced by other means, such as the method based on "scoring rules" adopted by de Finetti from the 1960s onwards. Unlike Ramsey and Savage, de Finetti does not establish a strict link between probability and utility, and regards probability as a subject characterized by an intrinsic value, to be dealt with autonomously.

De Finetti deems the foundation of probabilistic inference equally operational and grounds it on the so-called "representation theorem", resulting from a combination of Bayes' rule and exchangeability. A weaker property than independence, exchangeability has the advantage of allowing us to learn from experience more rapidly. De Finetti's great accomplishment was to show that the adoption of Bayesian method in conjunction with exchangeability leads to a convergence between prior probabilities and observed frequencies. Given that prior probabilities are for him the expression of degrees of belief, one can say that de Finetti's representation theorem shows how subjective probability can be applied to statistical inference, providing a model of how to proceed in such a way as to allow for an interplay between degrees of belief and frequencies. Like prior probabilities, de Finetti regards exchangeability as a subjective assumption, because its adoption depends on a personal choice.

As is well known, de Finetti strongly opposed the notion of "objective probability". The statement "probability does not exist", printed in capital letters in the "Preface" to the *Theory of Probability*[27], has fostered the impression that the subjective interpretation of probability is somewhat arbitrary. De Finetti's claim is inspired by the desire to keep the notion of probability free from metaphysical contaminations, and is rooted in his anti-realist and pragmatist philosophy of probability. It is noteworthy, however, that de Finetti struggles against *objectivism*, or the idea that probability depends entirely on some aspects of reality, not against *objectivity*. While opposing "the distortion" of "identifying objectivity and objectivism", deemed a "dangerous mirage" (de Finetti 1962, p. 344), he takes the problem of the objectivity of probability evaluations very seriously. Obviously, when a considerable amount of information on frequencies is available, it influences probability assignments through the assumption of exchangeability. Often, however, only scant information of this sort is available, in which case the problem of how to obtain good probability evaluations remains open. De Finetti's approach is based on *penalty methods*, like the so-called "Brier's

[27] English edition of de Finetti (1970).

rule", named after the meteorologist Brier, who applied it to weather forecasts. De Finetti did extensive work on scoring rules, partly in cooperation with Savage[28].

In order to grasp de Finetti's position it is important to recall his distinction between the *definition* and the *evaluation* of probability, which are seen as utterly different concepts that should not be conflated. De Finetti believes that the confusion between these two concepts imprints all the other interpretations of probability, whose upholders look for a unique criterion – be it frequency, or symmetries – and use it as grounds for both the definition and the evaluation of probability. By contrast, subjectivists take an "elastic" attitude according to which there are no "correct" probability assignments. As de Finetti put it: "the subjective theory [...] does not contend that the opinions about probability are uniquely determined and justifiable. Probability does not correspond to a self-proclaimed 'rational' belief, but to the effective personal belief of anyone" (de Finetti 1951, p. 218). This passage documents de Finetti's refusal of the "necessarist" attitude adopted by logicists like Jeffreys. Instead of committing the choice of one particular function to a single rule or method, de Finetti opts for a flexible attitude according to which such a choice is the result of a complex and largely context-dependent procedure involving both subjective and objective elements. In other words, objective elements do provide a basis for judgment, although not the only basis: "every probability evaluation essentially depends on two components: (1) the objective component, consisting of the evidence of known data and facts; and (2) the subjective component, consisting of the opinion concerning unknown facts based on known evidence" (de Finetti 1974, p. 7). Obviously, the degree to which the subjective component will influence probability evaluations is determined by the context; in a number of situations personal experience and expertise play a crucial role, whereas in others objective elements are preponderant. This should be the case of probability attributions grounded on scientific theories.

Although de Finetti never devoted specific attention to the use made of probability in science, committing himself to the conviction that science is just a continuation of everyday life and that subjective probability can do the whole job, in the last period of his production he entertained the idea that probabilities encountered in science derive a peculiar robustness from scientific theories. So, the posthumous volume *Filosofia della probabilità* contains a few remarks to the effect that probability distributions belonging to scientific theories – he refers specifically to statistical mechanics – provide "more solid grounds for subjective opinions" (de Finetti 1995, English edition 2008, p. 63). Unlike Ramsey and Jeffreys, however, de Finetti did not feel the need to include in his theory a notion of probability specifically devised for application in science. This is arguably a lacuna in his perspective, which is otherwise suggestive of important developments from both a technical and a philosophical point of view.

Like Reichenbach, de Finetti is an inductivist and a Bayesian, but unlike Reichenbach he denies that there are "unknown" correct probability values to be approached by repeated applications of the inductive procedure. For de Finetti the shift from initial to final probabilities means going from one subjective probability to another, and updating one's mind in view of new evidence does not mean changing opinion. In his words: "if we reason according to Bayes's theorem, we *do not* change our opinion. We keep the same opinion,

[28] For more on penalty methods see Dawid and Galavotti (2009). De Finetti's attitude towards the problem of objectivity is described in some detail in Galavotti (2001).

yet updated to the new situation. If yesterday I was saying 'It is Wednesday', today I would say 'It is Thursday'. However I have not changed my mind, for the day after Wednesday is indeed Thursday" (*ibid.*, p. 43). The idea of correcting previous opinions is totally alien to his perspective, as is the notion of a self-correcting procedure held by Reichenbach.

Like Reichenbach, de Finetti takes a pragmatist stand on the problem of induction, appealing to the notion of success. He applies this kind of argument to exchangeable functions, embedded in the Bayesian framework. By giving a mathematical argument – the representation theorem – showing the coherence of our inductive habits, de Finetti claims to offer a pragmatic solution to Hume's problem of induction, and to proceed one step further along the road Hume indicated, namely in the direction of constructing a purely empiricist theory of induction.

7.5 Concluding Remarks

The authors whose work has been reviewed here share the conviction that probability is an indispensable ingredient of science and knowledge at large, and that inductivism is an essential component of scientific method, although they disagree on the precise notion of probability to be put at its core. Reichenbach and de Finetti's option for a particular interpretation of probability is rooted in their different attitudes towards realism; the objectivist view of probability as frequency embraced by Reichenbach makes him a realist of sorts[29], whereas the subjectivist de Finetti is an extreme anti-realist[30]. Jeffreys' philosophy of probability seems to elude the polarization of realism and anti-realism, and his option for the logical interpretation of probability is motivated by the need to ensure probability an objective status while remaining in the realm of an epistemic view. A philosophical trait common to the perspectives of all three authors is pragmatism. This is far more evident in connection with Jeffreys' constructivism and de Finetti's operationalism, but also affects some aspects of Reichenbach's approach, such as the stress he puts on prediction, his justification of induction and his theory of meaning[31].

The authors considered here influenced the subsequent debate in complementary ways. Reichenbach has to be credited with the idea that causality can be taken in a probabilistic sense, and that understanding of the causal structure of the world can be grounded on it. Taken up by Wesley Salmon, who used it as a basis for his neo-mechanical theory of explanation, this idea gave rise to a flourishing research field. Although Jeffreys' influence is more confined to his work as a statistician, his "necessarist" attitude anticipated recent work in the foundations of probability, namely the current often called "objective Bayesianism". Strangely enough, Jeffreys' constructive epistemology has attracted little attention, in spite of its originality. De Finetti exerted a decisive influence both on statisticians and philosophers of science.

[29] See Salmon (1994) for a discussion of Reichenbach's realism.

[30] See Galavotti (1989) for a more detailed account of de Finetti's anti-realist philosophy of probability. See also Galavotti (2003b) for a comparison of Carnap, Reichenbach and de Finetti's views on probability.

[31] Some remarks on the influence of pragmatism on the debate on the foundations of probability are contained in Galavotti (2011b).

The probabilistic conception of epistemology pioneered by these authors has gradually become predominant among philosophers of science. Among those who foster this approach, special mention is due to Richard Jeffrey and Patrick Suppes. Jeffrey (1926–2002) has been the most earnest follower of de Finetti's radical probabilism, a position with which he identified himself. Jeffrey shares with de Finetti the tenet that the entire edifice of human knowledge rests on probability judgments, not on certainties, and embraces a form of Bayesianism deemed anti-rationalist, non-foundational, and pragmatist. In the first place, Jeffrey rejects as an "empiricist myth" the idea that experiential data can be grounded on a phenomenistic base, as well as the related claim that this kind of information forms the content of so-called "observation sentences", that can qualify as true or false. Such an assumption – typically held by logical empiricists including Carnap – combined with a Bayesian standpoint leads to a view according to which conditioning is based on experimental evidence taken as certain. Within Jeffrey's radical probabilism, this view is abandoned in favour of the conviction that probabilities need not be based on certainties: "it can be probabilities all the way down, to the roots" (Jeffrey 1992a, p. 11). In view of this, Jeffrey deems radical probabilism a "non-foundational methodology" (*ibid.*, p. 68), or a "softcore" version of empiricism (Jeffrey 2004, p. 59 ff.) because it does not ground knowledge on certainty and admits uncertain and incomplete evidence as a basis for conditioning.

Equally discarded is Carnap's idea that the empirical and rational components of our judgments can be sharply separated, together with the related tenet that conditioning starts from the assumption of some *a priori* distribution that can be justified on logical grounds. It is in this sense that Jeffrey qualifies his own probabilism as "anti-rationalist".

While departing from Carnap's programme Jeffrey turns to pragmatism, holding that judgments result from a multitude of factors, some of which are strictly pertinent to the evaluating subject, while others are essentially social products. In the process leading to the formulation of probability judgments, personal intuition and trained expertise mingle with a whole set of methods, theories, and skills shared by the scientific community. As Jeffrey puts it: "modes of judgments (probabilizing, etc.) and attendant standards of rationality are cultural artifacts, bodies of practice modified by discovery or invention of broad features seen as grounds to stand on" (Jeffrey 1992b, p. 203). Jeffrey's pragmatism, which closely resembles de Finetti's, does not contemplate a notion of rationality beyond the sphere of human knowledge and conduct to guide our judgment. While regarding this as the fundamental lesson that Bayesians should learn, Jeffrey warns that radical probabilism qualifies as a programme whose applicability to everyday practice cannot be canonized once and for all, but depends on context-dependent techniques forming "an art of judgment going beyond honest diligence" (*ibid.*, p. 204). At the core of Jeffrey's all-round probabilism lies the dynamic theory of decision that forms the content of *probability kinematics*, on which the author started working in the 1960s with his pioneering *The Logic of Decision*[32]. A key aspect of Jeffrey's perspective is the attempt to accommodate within his radical probabilism the notions of chance and objective probability[33]. This attitude is in line with the tendency, which has recently grown stronger, to open subjectivism to these notions.

[32] See also note 23.

[33] See Jeffrey (2004), pp. 19 ff.

Another vigorous supporter of probabilistic epistemology, Patrick Suppes (1922–2014), holds that "it is probabilistic rather than merely logical concepts that provide a rich enough framework to justify both our ordinary ways of thinking about the world and our scientific methods of investigation" (Suppes 1984, p. 2). On that premise, Suppes develops a view that he calls *probabilistic empiricism*, namely a probabilistic approach to science and philosophy intended to supersede the "neo-traditional metaphysics" centred on determinism, in the conviction that "certainty of knowledge – either in the sense of psychological immediacy, in the sense of logical truth, or in the sense of complete precision of measurement – is unachievable" (*ibid.*, p. 10).

Suppes calls himself a "semi-Bayesian" (Suppes 1983a, 1993, p. 214) to stress the non-orthodox character of his version of Bayesianism. For instance he argues in favour of randomization, usually rejected by orthodox Bayesians[34]. Suppes shows appreciation of Jeffreys' work, especially in connection with prior probabilities, and his writings abound with references to de Finetti. However, his position is in many ways more flexible than de Finetti's. A central feature of Suppes' perspective stems from the conviction that exact values should be substituted by probability intervals. In this spirit he spelled out, partly in collaboration with Mario Zanotti, a cluster of results on upper and lower probabilities, giving rise to an approach that they regard as a natural extension of de Finetti's line of thought[35]. While agreeing with de Finetti on a number of issues, including the tenet that probability and utility are notions to be dealt with separately, Suppes does not share de Finetti's uncompromising subjectivism. A major disagreement comes in connection with de Finetti's rejection of objective probability, a notion that is considered useful by Suppes, who makes the point that "there is not really an interesting and strong distinction between subjective and objective, or between belief and knowledge" (Suppes 1983b, p. 399). By contrast, the important thing is completeness of information, and complete knowledge in principle or in practice should be distinguished from incomplete knowledge with the possibility of learning more. When talking about the meaning of a probability statement, one should ask whether it is based on complete information, in the sense that there is no additional information that could bring about a change in the probability value. This is an important feature, especially when completeness of information arises from physical theory. In contexts where this kind of completeness obtains, probability acquires an objective meaning. Suppes believes that the notion of propensity can be a useful ingredient of the description of chance phenomena. More precisely, he claims that propensities confer an objective meaning upon the probabilities involved in the representation of a whole array of phenomena, such as coin tossing, radioactive decay, and many others, whose description involves structural considerations that put constraints on the estimation of probabilities. To be sure, Suppes does not take propensities to be probabilities, so his use of this notion has little to do with that made by upholders of the propensity theory of probability[36].

Suppes' epistemology is deeply pluralistic, in the sense that important concepts in science and philosophy – from theories to models of data, causal concepts, probability statements and inferential methods – are not given a univocal definition, being instead

[34] See Suppes (1983a).

[35] See the collection of their joint papers in Suppes and Zanotti (1996). The results on upper and lower probabilities by Suppes and Zanotti are summarized in Suppes (2002), chapter 5.

[36] See Suppes (2002), chapter 5 for more details on his view of propensities.

assigned a specific meaning depending on the context in which an instance occurs. Suppes' probabilistic and pragmatist epistemology, which the author started to develop in the 1950s, paved the way for subsequent research in a number of areas, among them the analysis of measurement and experimentation, probabilistic causality, procedural rationality, and the nature of models and their role in the making of scientific knowledge[37].

Jeffrey and Suppes' views are among the most far-reaching versions of probabilistic epistemology advanced in the literature. The work of authors such as Reichenbach, Jeffreys, Ramsey, and de Finetti paved the way for this kind of epistemology, which is increasingly gaining the favour of philosophers of science and scientists interested in foundational issues.

References

Bovens, Luc and Hartmann, Stephan (eds.) (2007) *Bayesian Epistemology*. Special issue of *Synthese*. 156. 3.

Carnap, Rudolf (1936–7) Testability and Meaning. *Philosophy of Science*. 3. pp. 419–71 and 4. pp. 1–40.

Carnap, Rudolf (1950) *Logical Foundations of Probability*. Chicago, IL: Chicago University Press. (2nd modified edition 1962. Reprinted 1967.)

Carnap, Rudolf (1952) *The Continuum of Inductive Methods*. Chicago, IL: Chicago University Press.

Carnap, Rudolf and Jeffrey, Richard (eds.) (1971) *Studies in Inductive Logic and Probability*. Vol. I. Berkeley-Los Angeles-London: University of California Press.

Cook, Alan (1990) Sir Harold Jeffreys. In *Biographical Memoirs of Fellows of the Royal Society*. 36. pp. 303–33.

Dawid, Philip A. and Galavotti, Maria Carla (2009) De Finetti's Subjectivism, Objective Probability, and the Empirical Validation of Probability Assessments. In Galavotti, M. C. (ed.) *Bruno de Finetti, Radical Probabilist*. pp. 97–114. London: College Publications.

de Finetti, Bruno (1931) Probabilismo. *Logos*, pp. 163–219. (English edition: Probabilism, *Erkenntnis* 31 (1989). pp. 169–223.)

de Finetti, Bruno (1937) La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*. 7. pp. 1–68. (English edition: Foresight: its Logical Laws, its Subjective Sources. In Kyburg, H. and Smokler, H. (eds.) (1964) *Studies in Subjective Probability*. pp. 95–158. New York-London-Sydney: Wiley. Second modified edition (1980). pp. 53–118. Huntington, NY: Krieger.)

de Finetti, Bruno (1951) Recent Suggestions for the Reconciliation of Theories of Probability. In Neyman, J. (ed.) *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. pp. 217–25. Berkeley, CA: University of California Press.

de Finetti, Bruno (1962) Does it Make Sense to Speak of 'Good Probability Appraisers'?. In Good, I. J. et al. (eds.) *The Scientist Speculates. An Anthology of Partly-Baked Ideas*. pp. 357–64. New York, NY: Basic Books.

de Finetti, Bruno (1968) Probability: the Subjectivistic Approach. In Klibansky, R. (ed.) *La philosophie contemporaine*. pp. 45–53. Florence: La Nuova Italia.

[37] A broad account of Suppes' contribution to philosophy of science is contained in Galavotti (1994). See also the other papers contained in the collection in three volumes in Humphreys (ed.) (1994), with Suppes' "Replies".

de Finetti, Bruno (1970) *Teoria delle probabilità*. Torino: Einaudi. (English edition: (1975) *Theory of Probability*. New York, NY: Wiley.)
de Finetti, Bruno (1974) The Value of Studying Subjective Evaluations of Probability. In Staël von Holstein, C.-A. (ed.) *The Concept of Probability in Psychological Experiments*. pp. 1–14. Dordrecht-Boston, MA: Reidel.
de Finetti, Bruno (1982) Probability and my Life. In Gani, J. (ed.) *The Making of Statisticians*. pp. 4–12. New York: Springer.
de Finetti, Bruno (1995) *Filosofia della probabilità*. Milan: Il Saggiatore. (English edition: (2008) *Philosophical Lectures on Probability* ed. by A. Mura. Dordrecht: Springer.)
Eberhardt, Frederick and Glymour, Clark (2011) Hans Reichenbach's Probability Logic. In Gabbay, D., Hartmann, S., and Woods, J. (eds.) *Handbook of the History of Logic X: Inductive Logic*. pp. 357–89. Amsterdam: Elsevier.
Galavotti, Maria Carla (1989) Anti-realism in the Philosophy of Probability: Bruno de Finetti's Subjectivism. *Erkenntnis*. 31. pp. 239–61.
Galavotti, Maria Carla (1991) The Notion of Subjective Probability in the Work of Ramsey and de Finetti. *Theoria*. 57. pp. 239–59.
Galavotti, Maria Carla (1994) Some Observations on Patrick Suppes' Philosophy of Science. In Humphreys (ed.) (1994). Vol. III. pp. 245–64.
Galavotti, Maria Carla (1995) F. P. Ramsey and the Notion of 'Chance'. In Hintikka, J. and Puhl, K. (eds.) *The British Tradition in the 20*th *Century Philosophy (Proceedings of the 17*th *International Wittgenstein Symposium)*. pp. 330–40. Vienna: Hölder-Pichler-Tempsky.
Galavotti, Maria Carla (1999) Some Remarks on Objective Chance (F. P. Ramsey, K. R. Popper and N. R. Campbell). In Dalla Chiara, M. L., Giuntini, R., and Laudisa, F. (eds.) *Language, Quantum, Music*. pp. 73–82. Dordrecht–Boston, MA: Kluwer.
Galavotti, Maria Carla (2001) Subjectivism, Objectivism and Objectivity in Bruno de Finetti's Bayesianism. In Corfield, D. and Williamson, J. (eds.) *Foundations of Bayesianism*. pp. 161–74. Dordrecht–Boston, MA: Kluwer.
Galavotti, Maria Carla (2003a) Harold Jeffreys' Probabilistic Epistemology: Between Logicism and Subjectivism. *British Journal for the Philosophy of Science*. 54. pp. 43–57.
Galavotti, Maria Carla (2003b) Kinds of Probabilism. In Parrini, P., Salmon, W. C., and Salmon, M. H. (eds.) *Logical Empiricism: Historical and Contemporary Perspectives*. pp. 281–303. Pittsburgh, PA: The University of Pittsburgh Press.
Galavotti, Maria Carla (2005) *Philosophical Introduction to Probability*. Stanford, CA: CSLI Publications.
Galavotti, Maria Carla (2011a) On Hans Reichenbach's Inductivism. *Synthese*. 181. pp. 95–111.
Galavotti, Maria Carla (2011b) Probability and Pragmatism. In Dieks, D., Gonzalez, W. J., Hartmann, S., Uebel, T., and Weber, M. (eds.) *Explanation, Prediction, and Confirmation.* pp. 499–510. Dordrecht: Springer.
Gillies, Donald (2000) *Philosophical Theories of Probability*. London–New York, NY: Routledge.
Good, Irving John (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.
Howie, David (2002) *Interpreting Probability*. Cambridge: Cambridge University Press.
Howson, Colin (1988) On the Consistency of Jeffreys's Simplicity Postulate, and its Role in Bayesian Inference. *The Philosophical Quarterly*. 38. pp. 68–83.

Humphreys, Paul (ed.) (1994) *Patrick Suppes: A Mathematical Philosopher*. 3 volumes. Boston, MA: Kluwer.

Jeffrey, Richard (1965) *The Logic of Decision*. Chicago, IL: The University of Chicago Press. (2nd edition 1983.)

Jeffrey, Richard (ed.) (1980) *Studies in Inductive Logic and Probability*. Vol. II. Berkeley-Los Angeles-London: University of California Press.

Jeffrey, Richard (1992a) *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Jeffrey, Richard (1992b) Radical Probabilism (Prospectus for a User's Manual). In Villanueva, E. (ed.) *Rationality in Epistemology*. pp. 193–204. Atascadero, CA: Ridgeview.

Jeffrey, Richard (2004) *Subjective Probability: The Real Thing*. Cambridge: Cambridge University Press.

Jeffreys, Harold (1931) *Scientific Inference*. Cambridge: Cambridge University Press. (Reprinted with Addenda 1937. 2nd modified edition 1957, 1973.)

Jeffreys, Harold (1933) Probability, Statistics and the Theory of Errors. *Proceedings of the Royal Society*. A. 140. pp. 523–35.

Jeffreys, Harold (1934) Probability and Scientific Method. *Proceedings of the Royal Society*. A. 146. pp. 9–16.

Jeffreys, Harold (1936) The Problem of Inference. *Mind*. 45. pp. 324–33.

Jeffreys, Harold (1937) Scientific Method, Causality, and Reality. *Proceedings of the Aristotelian Society*. New Series. 37. pp. 61–70.

Jeffreys, Harold (1939) *Theory of Probability*. Oxford: Clarendon Press. (2nd modified edition 1948, 1961, 1983.)

Jeffreys, Harold (1955) The Present Position in Probability Theory. *The British Journal for the Philosophy of Science*. 5. pp. 275–89. (Also in Jeffreys, H. and Swirles, B. (eds.) (1971–7). Vol. VI. pp. 421–35.)

Jeffreys, Harold (1963) Review of *The Foundations of Statistical Inference* by L. J. Savage. *Technometrics*. 5. pp. 406–10.

Jeffreys, Harold and Swirles, Bertha (eds.) (1971–7) *Collected Papers of Sir Harold Jeffreys on Geophysics and Other Sciences*. 6 volumes. London-Paris-New York: Gordon and Breach Science Publishers.

Lindley, Dennis (1991) Sir Harold Jeffreys. *Chance*. 4. pp. 10–21.

Martin-Löf, Per (1969) The Literature on von Mises' Kollektivs Revisited. *Theoria*. 35. pp. 12–37.

Popper, Karl Raimund (1959) The Propensity Interpretation of Probability. *British Journal for the Philosophy of Science*. 10. pp. 25–42.

Popper, Karl Raimund (1990) *A World of Propensities*. Bristol: Thoemmes.

Ramsey, Frank Plumpton (1931) *The Foundations of Mathematics and Other Logical Essays*, ed. by Braithwaite, R. B. London: Routledge and Kegan Paul.

Ramsey, Frank Plumpton (1990) *Philosophical Papers* ed. by Mellor, H. Cambridge: Cambridge University Press.

Reichenbach, Hans (1925) Die Kausalstruktur der Welt und der Unterschied zwischen Vergangenheit und Zukunft. *Sitzungsberichte, Bayerische Akademie der Wissenschaft*. Nov. pp. 133–75. (English edition: The Causal Structure of the World and the Difference between Past and Future.) In Reichenbach (1978). Vol. 2. pp. 81–119.

Reichenbach, Hans (1930) Kausalität und Wahrscheinlichkeit. *Erkenntnis*. 1. pp. 158–88. (English edition: (1959) Causality and Probability. In *Modern Philosophy of Science* ed. by

Reichenbach, M. pp. 67–78. London: Routledge and Kegan Paul. Reprinted in Reichenbach (1978) Vol. 2. pp. 333–44.)

Reichenbach, Hans (1935) *Wahrscheinlichkeitslehre*. Leyden: Sijthoff. (English expanded edition: (1949) *The Theory of Probability*. Berkeley-Los Angeles, CA: University of California Press. 2nd ed. 1971.)

Reichenbach, Hans (1936) Logicist Empiricism in Germany and the Present State of its Problems. *The Journal of Philosophy*. 6. pp. 141–61.

Reichenbach, Hans (1937) La philosophie scientifique: une esquisse de ses traits principaux. In *Travaux du IX Congrès International de Philosophie*. pp. 86–91. Paris: Hermann.

Reichenbach, Hans (1938) *Experience and Prediction*. Chicago-London: The University of Chicago Press.

Reichenbach, Hans (1951) The Verifiability Theory of Meaning. *Proceedings of the American Academy of Arts and Sciences*. 53. pp. 46–60.

Reichenbach, Hans (1956) *The Direction of Time*. Berkeley-Los Angeles. CA: University of California Press.

Reichenbach, Hans (1978) *Selected Writings, 1909–1953*, ed. by Reichenbach, M. and Cohen R. S. 2 volumes. Dordrecht-Boston: Reidel.

Salmon, Wesley C. (1953) The Uniformity of Nature. *Philosophy and Phenomenological Research*. 14. pp. 39–48.

Salmon, Wesley C. (1971) *Statistical Explanation*. In Salmon, W. C., Jeffrey, R., and Greeno, J. G. *Statistical Explanation and Statistical Relevance*. Pittsburgh, PA: University of Pittsburgh Press.

Salmon, Wesley C. (1984) *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Salmon, Wesley C. (1994) Carnap, Hempel and Reichenbach on Scientific Realism. In Salmon, W. C. and Wolters, G. (eds.) *Logic, Language, and the Structure of Scientific Theories*. pp. 237–54. Pittsburgh-Konstanz: University of Pittsburgh Press-Universitätsverlag.

Salmon, Wesley C. (1998) *Causality and Explanation*. New York-Oxford: Oxford University Press.

Suppes, Patrick (1983a) Arguments for Randomizing. In Asquith, P. and Nickles, T. (eds.) *PSA 1982*. pp. 464–75 (Reprinted in Suppes (1993). pp. 213–24.)

Suppes, Patrick (1983b) The Meaning of Probability Statements. *Erkenntnis*. 19. pp. 397–403.

Suppes, Patrick (1984) *Probabilistic Metaphysics*. Oxford: Blackwell.

Suppes, Patrick (1993) *Models and Methods in the Philosophy of Science: Selected Essays*. Dordrecht-Boston: Kluwer.

Suppes, Patrick (2002) *Representations and Invariance of Scientific Structures*. Stanford, CA: CSLI Publications.

Suppes, Patrick and Zanotti, Mario (1996) *Foundations of Probability with Applications*. Cambridge-New York: Cambridge University Press.

von Mises, Richard (1928) *Wahrscheinlichkeit, Statistik und Wahrheit*. Vienna: Springer. (English edition (1939) *Probability, Statistics and Truth*. London-New York: Allen and Unwin. Reprinted (1957) New York: Dover.)

Williamson, Jon (2009) Philosophies of Probability. In Irvine, A. (ed.) *Handbook of the Philosophy of Mathematics*. Volume IX of the *Handbook of the Philosophy of Science*. pp. 493–533. Amsterdam: Elsevier.

Zabell, Sandy (1982) W. E. Johnson's 'Sufficientness' Postulate. *The Annals of Statistics*. 10. pp. 1091–9.

Zellner, Arnold (1982) Is Jeffreys a 'Necessarist'? *The American Statistician*. 36. pp. 28–30.

PART II

FORMALISM

CHAPTER 8

KOLMOGOROV'S AXIOMATIZATION AND ITS DISCONTENTS

AIDAN LYON

"[... T]here is little disagreement about the truth of the theory—indeed, it would not be an exaggeration to say that the theory of probability is commonly regarded as though it were necessarily true." Humphreys (1985), p. 568.

8.1 Introduction

Discontent with Kolmogorov's axioms? Surely discontent with The Axioms is as insane as discontent with The Truths of Logic. Indeed, some have argued that Kolmogorov's axioms *are* logic, the logic of partial belief, a natural—and the only—generalisation of classical logic, the logic of full belief (e.g., Jaynes (2003)).

Of course, many philosophers and logicians have expressed much discontent with classical logic. Admittedly, many philosophers and logicians are perhaps not quite bastions of sanity; however, their reasons for revising classical logic are perfectly sane. There are significant issues with vagueness, indeterminacy, inconsistency, and the like that cause potential problems for classical logic. Any such revision of classical logic can underpin a corresponding revision of the probability axioms. Chapter 12, *Probability and Nonclassical Logic*, in this volume, focuses on such non–classical probability theories and reasons for adopting them. This chapter focuses mostly on other sources of discontent.

Even though Kolmogorov's axioms are the orthodox probability axioms, they appear to be incompatible with the most common so-called "interpretations" of probability. Finite actual frequencies, infinite hypothetical frequencies, propensities, degrees of entailment, and even rational partial belief all appear to fail to satisfy Kolmogorov's axiomatization of probability. Or, to put it from another perspective: Kolmogorov's axioms appear to fail to satisfy each of the most common theories of what probabilities are. Each of these failures is a source

of possible discontent. In addition to this, there are reasons for discontent that seem to be independent of any interpretation of probability, and there is reason for discontent from the sciences too.

This chapter will survey the most common reasons to be discontented with Kolmogorov's axiomatization. Along the way, I will consider possible responses that one might make in the axiomatization's defence. First, though, we should be clear as to what all the discontent is about.

8.2 Kolmogorov's Axiomatization of Probability

Let Ω be a non-empty set, over which $\mathscr{F}$ is an algebra. Let P be a function from $\mathscr{F}$ to $\mathbb{R}$. If P satisfies the axioms:

$$\text{(K1) } P(A) \geq 0$$

$$\text{(K2) } P(\Omega) = 1$$

$$\text{(K3) } P(A \cup B) = P(A) + P(B), \text{ if } A \cap B = \emptyset$$

for every A and B in $\mathscr{F}$, then P is a *probability function*, and $(\Omega, \mathscr{F}, P)$ is a *probability space*. These axioms are often called *non-negativity*, *normalization*, and *finite additivity*, respectively.

If $\mathscr{F}$ is a σ–algebra, finite additivity is extended to so called *countable addivity*:

$$\text{(K3')} \; P\left(\bigcup_{i=1}^{i=\infty} A_i\right) = \sum_{i=1}^{i=\infty} P(A_i),$$

where the A_i are mutually disjoint.

Kolmogorov then adds a definition of conditional probability to the above axioms:

$$\text{(CP) } P(A|B) = \frac{P(A \cap B)}{P(B)}, \text{ where } P(B) > 0$$

for every A and B in $\mathscr{F}$. I shall call K1–3' + CP *Kolmogorov's axiomatization of probability*.

Before we move on to reasons why one might be unhappy with Kolmogorov's axiomatization, it will be worthwhile to pause for a moment, and think about the relation between the formal theory of probability and the concept of probability.

8.3 Where to Place the Blame?

Kolmogorov's axiomatization is quite remarkable in many ways. It situates probability theory as a branch of a more general mathematical theory, the theory of measures. And its

three simple axioms have resulted in numerous remarkable theorems, which have countless important applications throughout the sciences. It is simultaneously incredibly simple, unified, and powerful. And yet, there are reasons to be dissatisfied with it nonetheless.

It seems, *prima facie*, strange to be discontented with a piece of mathematics. One can be discontented with all sorts of things—politicians, the dinner menu, software, society, etc.—but these are all things that are changeable in some sense. The facts of mathematics, on the other hand, are apparently unchangeable. Being discontented with a piece of mathematics would therefore seem to be on a par with wishing that "up" "down".

However, discontent with mathematics can produce and has led to significant progress. Mathematicians have expressed discontent with their theories for a variety of reasons. They sought the Zermelo–Fraenkel axioms to avoid the inconsistency of naïve set theory. And then there was—and still is—the Axiom of Choice controversy, with some arguing we should accept it based on its necessity for several important theorems, and others arguing that we shouldn't because it leads to counterintuitive objects or that it isn't acceptable on constructivist grounds.

Such expressions of discontent arise when the truth of the relevant mathematics is clearly in question. Inconsistent set theory cannot possibly be true (according to classical logic), and there was *a priori* epistemic uncertainty surrounding the Axiom of Choice. Sometimes the uncertainty surrounding the mathematics in question has an empirical or "external" flavour to it. Perhaps the clearest example of this comes from geometry. Initially, the Parallel Postulate was called into doubt, as an axiom, because of its lack of self-evidence. Moreover, once it was realised that the postulate is false for the actual physical space that we happen to inhabit, it was quickly rejected, and so-called non-Euclidean geometries quickly flourished.

As we will see, the reasons for discontent with Kolmogorov's axiomatization are often external in this way—i.e., they come from particular applications of the theory. Applications include capturing the epistemic norms of graded belief, the behaviour of chance, and statistical reasoning. It is when the axiomatiziation is used in these applications that problems arise and discontent may ensue. Just as the truth of the Parallel Postulate for our actual physical space was called into question, we can ask if, e.g., countable additivity is true for the epistemic norms of people's degrees of belief, or if the definition of conditional probability is true for conditional physical chance, etc.

In any particular application of Kolmogorov's axiomatization, it is important to keep in mind what exactly the link is between the formalism and the target of the formalism. I began this chapter somewhat dramatically, with an analogy between probability and logic. This is a useful analogy to keep in mind as we consider various sources of discontent with Kolmogorov's axiomatization. As already mentioned, the sources of discontent often stem from particular applications of the axioms. In any such application, there has to be some link between the purely abstract axioms and the target of the application. For example, we will often want P to represent an agent's *degrees of belief*. Call such a link a *bridge principle*. Often, when things go wrong, there are two places where the blame can be laid: with the bridge principle in question or with the axiomatization.

Compare this with a potential source of discontent with classical logic. According to classical logic, $A \wedge \neg A$ entails everything—this is known as *explosion*. Why is explosion bad? Suppose a person has inconsistent beliefs. If classical entailment models the normative closure of belief, then that person ought to believe everything. Some authors have concluded from this absurdity (along with other considerations) that we need to revise classical logic,

and adopt some non-classical logic, in which explosion does not hold (e.g., Meyer (1971), pp. 814–15). However, as Harman has pointed out, we could equally well question the bridge principle involved: that classical entailment models the normative closure of belief (Harman (1986), p. 6).

Even when we're sure that the axioms deserve the blame, it may not be clear in a given case which axiom deserves that blame. No axiom is an island, entire of itself; the axioms work in tandem to produce the results of Kolmogorovian probability theory—whether those results are preferable or not. And so when trouble arises, more than just one axiom can often be blamed. Moreover, the blame can even be placed with something that the axioms presume—e.g., that probability values are real numbers, that the objects of probability are sets, that they form an algebra, etc.

Finally, there is a third place where the blame can be laid: with the application itself. As we'll see, some interpretations of probability have been criticized for not satisfying Kolmogorov's axiomatization. Is this a problem for the interpretation, or for the axiomatization—or for the bridge principle that links them? It's hard to say, but as we'll see, almost *every* interpretation has some degree of discomfort with Kolmogorov's axioms—even finite frequentism!

8.4 Discontent with Countable Additivity

Consider a fair lottery with denumerably many tickets. Since the lottery is fair, each ticket has equal probability of being drawn. But there are only two ways in which this can happen, and on both ways, trouble looms. On the first way, each ticket has some positive probability of winning, and any positive probability added to itself denumerably many times is larger than one, and so this is a violation of the axioms. On the second way, each ticket has zero probability of being drawn. Zero added to itself denumerably many times is zero, which is less than one, and so this also is a violation of the axioms. It's a case of too much or too little: on both ways of the lottery being fair, the probability that *some* ticket is drawn is either greater than one or less than one, and so there is a violation of the axioms. It seems, then, that Kolmogorov's axiomatization rules out—*a priori*—the possibility of a fair, countably infinite lottery. (Incidentally, the problem is reminiscent of Zeno's Paradox of Plurality. See, e.g., Salmon (2001), pp. 13–15.)

What's causing the problem? The argument—originally due to de Finetti (1974)—is typically used as an objection to K3', the axiom of countable additivity. However, one might also use it to object to K2, the axiom of normalization, which is clearly involved in the generation of the problem. Both of these options would be to blame the axioms. One can also try to blame the application. Some have argued that there is no problem here because there is no physical device that could set up such chances (Spielman (1977), Howson and Urbach (1989)). However, a response to this is that it doesn't matter whether such a physical device exists or not; a rational agent should be allowed to assign equal credence to each ticket winning (Williamson (1999)).

Can one blame the bridge principle? Yes. To set up the problem in a complete and precise way, one typically has an algebra that includes denumerably many events, one of each corresponding to each of the denumerably many tickets winning. One can then prove that there is no uniform probability distribution over those events. An alternative way to approach the problem, and so an alternative bridge principle, would be to start with a finite algebra, and a uniform distribution (which is unproblematic), and let the algebra increase to any arbitrary-but-finite-size (see, e.g., Jaynes (2003)). This is an unconventional line of thought, but it shows that there is a way to question the bridge principle involved, i.e., how one should model the situation (whether it be chances or someone's ideal credences) in the mathematics.

Another way to blame the bridge principle is to say that, in this particular situation, there is *no* bridge principle between our credences and Kolmogorov's axioms. As Bartha (2004) points out, one may argue that we don't have real-valued degrees of belief in this situation (pp. 306–7). If, instead, we only have relative probabilities, then one can maintain that the axiom of countable additivity is true because it doesn't apply in the case of de Finetti's lottery (pp. 309–10). (See Bartha (2004) for a more detailed discussion and constraints on relative probabilities.)

8.5 Discontent with Finite Additivity

Although on much safer ground, finite additivity has its discontents too. The structure it imposes on probability makes it difficult for probability to represent states of pure ignorance.

Consider a situation in which we are reasoning about whether some event, E, will occur or not. If we know nothing about E, we have no evidence for or against it, we have no convictions or intuitions about it, then it it seems to be unreasonable to presume E to be more likely than $\neg E$, or vice versa. So if we are to assign probabilities to E and $\neg E$, we have to give them *equal* probabilities. It then follows, from finite additivity, that they both have to be given a probability of $1/2$.[1] Compare this with a situation in which we know lots about E and $\neg E$—e.g., they are the outcomes of a fair flip of your lucky coin. In both situations, the probability calculus represents you as equally confident of E. And yet this seems not to be so.

Worse still is that the above reasoning can lead to paradox. If a cube factory produces cubes and those cubes can have side lengths between 0 and 1m, what is your probability that the next cube has its side length between 0 and 1/2m? Either it has this side length, E, or it does not, $\neg E$. By the reasoning above, the probability of E is therefore $1/2$. But if the cube factory produces cubes and those cubes have face areas between 0 and 1m^2, what is your probability that the next cube has its face area between 0 and $1/4\text{m}^2$? Here the reasoning is slightly different, but similar enough. It seems there are four events for which we cannot suppose any to be more probable than any other: the face area is between (i) 0 and 1/4, (ii) 1/4 and 1/2, (iii) 1/2 and 3/4, and (iv) 3/4 and 1. If these all have equal probability, then, by

[1] From equal probabilities, we have $P(E) = P(\neg E)$; from finite additivity we have $P(E) + P(\neg E) = 1$; and putting these together we get: $2P(E) = 1$, $2P(\neg E) = 1$, and so $P(E) = 1/2$ and $P(\neg E) = 1/2$.

finite additivity, they all have probability of 1/4. But now we have a contradiction, for $E \equiv$ (i) (a side length of 1/2 is the same as a face area of 1/4), and yet we have given them different probabilities. This example is taken from van Fraassen (1989), and is representative of a larger class of paradoxes—see also, e.g., Keynes (1921) and Jaynes (1973).

As usual, there are many places where the blame for these problems can be laid, and one place that can be blamed is finite additivity. If we drop finite additivity, then from $P(E) = P(\neg E)$ it does not follow that $P(E) = P(\neg E) = 1/2$. It is now permissible that, say, $P(E) = P(\neg E) = 0$. And perhaps that is the right probability assignment in a case of total ignorance. If probability is credence and you have no evidence or prior knowledge, then there is nothing to lend any credence to E or $\neg E$; you therefore have no credence in E or $\neg E$. Assigning vacuous probabilities to all of the possibilities for which you are completely ignorant avoids the above contradiction, and it also distinguishes you from the person who is well-informed about the options, but happens to have uniform credences.

One popular theory that drops finite additivity is the Dempster–Shafer theory, sometimes called the theory of *belief functions* (see, e.g., Shafer (1976)). See also Ghirardato (2001) for reasons for adopting non-additive probabilities owing to cases of ignorance.

8.6 Discontent with Conditional Probability

According to Kolmogorov's definition of conditional probability, $P(A|B)$ is undefined whenever $P(B) = 0$. As many authors have noted, there are countless situations where this appears to be false.

Suppose we choose a point randomly on the surface of the Earth, where we assume the Earth is a perfect sphere. What is the probability that the point is in the Western Hemisphere, given that it is on the Equator? The answer is surely 1/2, and yet according to Kolmogorov, it is undefined, because the probability of the point's being on the Equator is 0. (In this example, unconditional probability corresponds to relative area of the surface of the Earth. The event of the point being on the Equator gets a probability of 0 because the Equator is a line and therefore has area of 0.)

What's the cause of this problem? The natural response is that the problem is due to the overly restrictive definition of conditional probability. Indeed, Kolmogorov himself was fully aware of this sort of problem and developed more general, and complicated, definition of conditional probability (Kolmogorov (1933/1950). However, this complicated definition of conditional probability runs into problems of its own (see, e.g., Seidenfeld et al. (2001) and Hájek (2003)).

Other authors have developed alternative axiomatizations that take conditional probability as primitive and define unconditional probability in terms of it (e.g., Rényi (1955), Popper (1959a)). Such axiom systems allow $P(A|B)$ to be given a definite value even if $P(B|\Omega) = 0$, which in these contexts is typically understood as the unconditional probability of B.

Another common response to the problem is to argue that in cases such as the Equator example, $P(B)$ is not really equal to 0. This response places the blame with the assumption that probabilities are real–valued, and so I'll postpone discussion of it until section 8.7, where I discuss discontent with this assumption more generally.

Also according to Kolmogorov's definition of conditional probability, $P(A|B)$ is undefined whenever the unconditional probabilities in which it is defined are undefined. What is the probability that Joe gets "heads", given he flips a fair coin? Answer: 1/2. But what is the probability that Joe flips a fair coin? Answer: Who knows?! Whether Joe flips the coin may be a matter of free will, in which case there may be no chance associated with it (Hájek (2003)). Never mind matters of free will, there just may be no unconditional probability of Joe's flipping the coin. There may be a *conditional* probability of Joe's flipping *given* that someone asks Joe nicely to flip it, or pays him, or if he is in a coin flipping mood, or if he has sworn to never flip coins again, etc. But a free-floating unconditional probability of Joe's flipping the coin? It's not at all clear that such a probability exists; and even if it does, if by probability we mean rationally required degree of belief, then there may nevertheless be no such thing anyway. It seems you can have a conditional degree of belief of 1/2 that the coin lands "heads", given that Joe flips the coin fairly, and not be rationally required to have some degree of belief that Joe flips the coin fairly.

The problem seems to be due to Kolmogorov's definition of conditional probability in terms of unconditional probability. Perhaps the most natural response, then, is that we should axiomatize conditional probability directly, instead of defining it in terms of unconditional probability. This approach also seems to have the advantage of also solving the problem from the Equator example. Hájek, for example, concludes that we should move to an alternative axiom system that takes conditional probability as primitive (Hájek (2003), pp. 315–16). Unfortunately, many (if not all) such alternative axiom systems don't really solve the problem. Many of the alternative axiom systems require that probabilities such as P(I flip a fair coin | I flip a fair coin or I don't) be defined. But for the same reasons as before, it seems there is intuitively no such probability (see Lyon (2014) for a discussion of this point in context of the propensity interpretation of probability). The problem is therefore not Kolmogorov's alone and it seems that something more drastic has to be done.

So far I have been focusing on problems where the conditional probabilities are undefined when they should be defined. But the opposite also happens. Just as problematic are cases where conditional probabilities are defined when they should be undefined, or defined in some other way. For example, according to the propensity interpretation of probability, $P(E|C)$ is the propensity of C to produce, or cause, E. However, Humphreys (1985) has shown that in many situations Kolmogorov's axiomatization requires that if $P(E|C)$ has a value, then so does $P(C|E)$. This seems strange, since presumably causes must precede effects, and yet Kolmogorov's axioms seem to tell us that an effect can have a propensity to produce its cause. Surely that is wrong: the lighting of a match in the evening has a propensity to burn down the factory at night, but the burning down of the factory at night doesn't have a propensity to light the match in the evening. This problem for Kolmogorov's understanding of conditional probability for the propensity interpretation is known as Humphreys' Paradox. The standard response to this problem is that it is trouble for the propensity interpretation—i.e., the application gets the blame. Humphreys' own conclusion, however, is different: he blames the axiom system (Humphreys (1985), pp. 568–9).

8.7 Discontent with Positive Real Numbers

When faced with undefined conditional probabilities or infinite lotteries, there is yet a another aspect of the axioms that we could blame: the assumption that probabilities are *real numbers*.

In both situations, we want a probability value that is really small, but not 0. The problem with the real numbers is that there is no number that is small enough but not so small that it is 0. This is easiest to see in de Finetti's lottery example: as soon as the probability of each ticket's winning is greater than 0, then no matter how small that number is, the total probability will sum to more than 1. One response, then, is to give up on the real number system, and move to a richer system that has numbers that are smaller than every real number, but that are greater than 0. Such numbers are called *infinitesimals*, and the most cited number system that contains them is the system of hyperreals (e.g., Robinson (1966)).

If probability values are hyperreals, then we can say that each ticket in de Finetti's lottery has some infinitesimal probability of winning (e.g., Bartha and Hitchcock (1999)). Similarly, we can say that the event of choosing a point on the Equator has infinitesimal probability, and so the conditional probability of choosing a point in the Western Hemisphere given the chosen point is on the Equator doesn't go undefined (e.g., Lewis (1980), pp. 267–8). It would seem that infinitesimals provide an elegant solution to two different problems.

In fact, there is even more good news for infinitesimal fans. It seems to be a dictum of rationality that one should be open-minded: one should not assign zero probability to any event that one considers to be possible. This principle often goes by the name of *Regularity*. According to Kolmogorov's axioms, though, it is mathematically impossible to satisfy this principle when dealing with uncountably many events (e.g., Hájek (2003), pp. 281–2). One is forced to assign zero probability to uncountably many events that one considers to be possible. However, this is not true if probability values can take on infinitesimal values.

So infinitesimals seem to solve three distinct sources of discontent with Kolmogorov's axioms. (Moreover, there is a sense in which adopting infinitesimals does not amount to a revision of the axioms, for the hyperreal system is a non-standard model *of the reals*. One could read Kolmogorov's axioms as leaving the model of the reals as unspecified, in which case we don't really have a case of discontent with the axiomatization.) Infinitesimals have their problems, though. It's impossible to name one of them, for example. So when we say that each ticket in de Finetti's lottery has an infinitesimal probability of winning, *which* infinitesimal is it? It's strange that we can't answer that question (see Hájek (2003), pp. 292–3 for more details). More seriously, though, is that even if we allow probability values to be infinitesimals, we are *still* forced to assign zero probability to possible events (Williamson (2007)).

Instead of moving to a richer number system, some may want to move to a *poorer* number system. According to Kolmogorov's axiomatization, $P(A) = 1/\pi$ is a legitimate probability assignment—i.e., there are probability functions that make such an assignment. Those who believe in the finite actual frequency interpretation of probability have to disagree. Nothing can occur with a $1/\pi$ relative frequency in a finite number of trials, and so $P(A) = 1/\pi$ cannot be a legitimate probability assignment. Finite frequentists have to insist that P is a

function from $\mathscr{F}$ to $\mathbb{Q}$, and not $\mathbb{R}$. (Of course, this point is often used as an objection to finite frequentism (e.g., Hájek (1997), pp. 224–5)—but as usual, the sword can cut both ways.)

So far I have been discussing discontent with the assumption that probability values are real numbers. Kolmogorov's axiomatization says something stronger, though: the axiom of non-negativity, KP1, makes sure that they are *positive* real numbers. Why the ban on negative numbers? One reason is that it is not at all clear how one would interpret them: if $Pr(A) = -0.5$, should we expect A to happen about minus fifty percent of the time? (!)

Despite our inability to make sense of negative probabilities, they nevertheless appear in the sciences. They appear in quantum mechanics, and even in classical physics (see Muckenhaim et al. (1986)). They also make an appearance in financial mathematics (Haug (2007), Meissner and Burgin (2011)) and machine learning (Lowe (2007/2004)). Negative probabilities typically only appear as intermediate steps in calculations—much like $3 = 4 + (-1)$—or as probabilities of unobservables. However, see Burgin (2010) for a frequency interpretation of negative probabilities.

Another source of discontent with the probability values of Kolmogorov's axiomatization is that it assumes that probabilities are point values. Many authors have argued for interval-valued probabilities. For example, it seems rationally permissible to not have a point-valued credence in all propositions. Does your confidence that it will rain tomorrow have to be precise to infinitely many decimal points for you to be rational? Arguably, no it doesn't. Instead of someone having precise credences, an agent may have lower and upper bounds on their credences—forming so-called *indeterminate credences*. See, e.g., Walley (1991), Kyburg and Pittarelli (1996), Levi (2000), and Hájek and Smithson (2012) for more details.

One final reason to be discontented with the probability values of Kolmogorov's axiomatization is with the very assumption that there are probability *values*. Fine (1973) argues that a comparative probability axiom system (i.e., one that axiomatizes "A is more probable than B") is more general and powerful than Kolmogorov's axiom system.

8.8 Discontent with Sets and Algebras

According to Kolmogorov's axiomatization the objects of probabilities are sets. This immediately rules out any interpretation of probability that thinks the objects are sentences, or events, or even propositions, if propositions are not sets of something (e.g., worlds). For these reasons, Popper thought that a formal theory of probability should make no assumption regarding what the bearers of probability are:

> In Kolmogorov's approach it is assumed that the objects a and b in $p(a,b)$ are sets (or aggregates). But this assumption is not shared by all interpretations: some interpret a and b as states of affairs, or as properties, or as events, or as statements, or as sentences. In view of this fact, I felt that in a formal development, no assumption concerning the nature of the 'objects' or 'elements' a and b should be made [...]. Popper (1959b), p. 40.

Popper goes on to develop an axiom system that makes no such assumptions and has the virtue—in Popper's eyes—that the algebraic structure of the objects of probability emerges as a consequence from the axioms, rather than being built into them (Popper (1959a)).

However, some have argued that the objects of probability should not necessarily form an algebra, or a σ–algebra, so this property should be neither built into the axioms nor a consequence of them. For example, consider the following:

> A class of photographs may be such that the probability of a predominantly dark pictures is $\frac{1}{2}$ and the probability of a grainy texture is also $\frac{1}{2}$. However, we may have no data on, and little interest in, whether grainy patterns tend to be dark or not. Should we then be prevented from using probability theory to model this sample source when designing pattern classifiers for this problem? Fine (1973), p. 62.

Fine's point is that there are situations where we can assign probabilities to A and B but not to their union $A \cup B$ or intersection $A \cap B$. And yet Kolmogorov's axioms, because they require probabilities to be defined over an algebra, require these probabilities to be defined.

One alternative to the regular algebra is the Λ–*field*. The closure properties of a Λ–field allow that A and $B \in \Lambda$, but $A \cap B \notin \Lambda$. An example would be $\Lambda = \{\emptyset, \Omega, (a,b), (c,d), (a,d), (b,c)\}$, which could be interpreted as:

> A photograph is
> (a) dark colour and grainy texture
> (b) dark colour and smooth texture
> (c) light colour and smooth texture
> (d) light colour and grainy texture

Note that the event of a photograph being dark and grainy is not in Λ, and so it doesn't get a probability (Fine (1973), p. 63). The neat thing about using a Λ–field is that it only requires the introduction of events whose probabilities can be calculated from the given probabilities using K1–K3. A downside is that it doesn't allow us to define many conditional probabilities that seemingly ought to be defined—e.g., we can't say that $P(\text{dark colour}|\text{dark colour and grainy texture}) = 1$.

Another alternative to the σ–field is the von Mises field, the $\mathscr{V}$*–field*. The precise definition of the $\mathscr{V}$–field is complicated and the interested reader should see, e.g., Fine (1973), pp. 64–5 for a definition. One important difference between the $\mathscr{V}$–field and the σ–field is that the former contains only events that can be measured using finite data sets. For example, for an indefinitely long series of coin flips, the event "heads occurs only finitely often" does not appear in the $\mathscr{V}$–field (*ibid*). Another important difference is that the set of events with hypothetical limiting frequencies form a $\mathscr{V}$–field but not a σ–field (nor even a field). (Von Mises was a hypothetical frequentist.)

8.9 Conclusion

Who could be discontented with Kolmogorov's axiomatization of probability? The answer seems to be: almost everyone! We have seen reasons to be discontented with almost every aspect of Kolmogorov's axiomatization: that the objects of probabilities are *sets*, and that

they form an *algebra* of sets; that probability values are *numbers*, that they are *real* numbers, that they are *positive* real numbers, that they have an *upper bound*, and that there are even probability *values*; that probabilities are *countably additive*, and even that they are *additive*; and finally that conditional probabilities are *ratios of unconditional probabilities*.

With so much discontent, how has Kolmogorov's axiomatization achieved its level of orthodoxy? Perhaps what lies behind it is that the axiomatization is the best compromise among many competing demands on the formal theory of probability. It's a wonder that it has such a wide domain of applicability to a concept that we understand so little and use so widely. Probability is everywhere, the guide to life as some have (often) said, and indispensable to the sciences; and yet it is notoriously difficult to analyse. It's somewhat amazing, then, that there is a such a successful formal theory of something that is so conceptually messy.

Nevertheless, we might wonder why probability shouldn't go the way of geometry, with different axiomatizations appropriate for different applications. Put this way, Kolmogorov's axiomatization is the Euclidean geometry of probability. It's still useful for many purposes and often a good first approximation, but there are other axiomatizations that are better suited for particular tasks.

Acknowledgements

Many thanks to Alan Hájek for helpful comments on an earlier draft of this paper. This work was supported by the Humboldt Foundation, the Centre of Excellence for Biosecurity Risk Analysis, and the Australian National University (ARC grant number DP1097075).

References

Bartha, P. (2004) Countable Additivity and the de Finetti Lottery. *British Journal for the Philosophy of Science.* 55. pp. 301–21.

Bartha, P. and Hitchcock, C. (1999) The Shooting-Room Paradox and Conditionalising on "Measurably Challenged" Sets. *Synthese.* 118. pp. 403–37.

Burgin, M. (2010) Interpretations of negative probabilities. [Online] *Arxiv Preprint.* arXiv:1008.1287 [physics. data-an] Available from: arxiv.org/abs 1008.1287. [Accessed 12 Oct 2015.]

de Finetti, B. (1974) *Theory of Probability*. Vol. 1. Translated from the Italian by Machi and Smith. New York, NY: Wiley.

Fine, T. (1973) *Theories of Probability*. New York, NY: Academic Press.

Ghirardato, P. (2001) Coping with Ignorance: Unforeseen Contingencies and Non–Additive Uncertainty. *Economic Theory.* 17. pp. 247–76.

Hájek, A. (1997) 'Mises Redux'–Redux: Fifteen Arguments Against Finite Frequentism. *Erkenntnis.* 45. pp. 209–27.

Hájek, A. (2003) What Conditional Probability Could Not Be. *Synthese.* 137. 3. pp. 273–323.

Hájek, A. and Smithson, M. (2012) Rationality and indeterminate probabilities. *Synthese.* 187 (1), pp. 33–48.

Harman, G. (1986) *Change in View: Principles of Reasoning*. Cambridge, MA: MIT Press.

Haug, E. G. (2007) *Derivatives Models on Models*. New York, NY: Wiley.

Howson, C. and Urbach, P. (1989) *Scientific Reasoning*. 1st edition. La Salle, IL: Open Court.
Humphreys, P. (1985) Why Propensities Cannot Be Probabilities. *Philosophical Review.* 94. pp. 557–70.
Jaynes, E. T. (1973) The Well Posed Problem. *Foundations of Physics.* 4. 3. pp. 477–92.
Jaynes, E. (2003) *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
Keynes, J. M. (1921) *A Treatise on Probability*. London: Macmillan.
Kolmogorov, A. N. (1933/1950) *Foundations of the Theory of Probability*. Translated from the German by Nathan Horison. Chelsea Publishing Company.
Kyburg H. E. Jr., and Pittarelli, M. (1996) Set-based bayesianism. *IEEE Transactions on Systems, Man and Cybernetics*. Part A: *Systems and Humans*. 26. 3. pp. 324–39.
Levi, I. (2000) Imprecise and Indeterminate Probabilities. *Risk, Decision and Policy.* 5. pp. 111–22.
Lewis, D. (1980) A Subjectivist's Guide to Objective Chance. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability*. Vol. II. Berkeley, CA: University of California Press.
Lowe, D. (2007) Machine Learning, Uncertain Information, and the Inevitability of Negative 'Probabilities'. [Online video] In *Machine Learning Workshop 2004*. Sheffield, England. Available from: videolectures.net/mlws04_lowe_mluii/. [Accessed 12 Aug 2015.]
Lyon, A. (2014) From Kolmogorov, to Popper, to Rényi: There's no escaping Humphreys' Paradox (When Generalised). In Wilson, A. (ed.) *Chance and Temporal Asymmetry*. Oxford: Oxford University Press.
Meissner, G. and Burgin, M. (2011) Negative Probabilities in Financial Modeling. [Online] Available from: papers.ssrn.com/sol3/paperscfm?abstract_id=1773077. [Accessed 12 Aug 2015.]
Meyer, R. (1971) Entailment. *The Journal of Philosophy.* pp. 808–18.
Muckenheim, W., Ludwig, G., Dewdney, C., Holland, P., Kyprianidis, A., Vigier, J., Petroni, N., Bartlett, M. and Jaynes, E. T. (1986) A Review of Extended Probability. *Physics Reports.* 133. pp. 337–401.
Popper, K. R. (1959a) *The Logic of Scientific Discovery*. New York: Basic Books.
Popper, K. R. (1959b) The Propensity Interpretation of Probability. *The British Journal for the Philosophy of Science.* 10. 37. pp. 25–42.
Rényi, A. (1955) On a New Axiomatic Theory of Probability. *Acta Mathematica Academiae Scientiarum Hungaricae.* 6. pp. 286–335.
Robinson, A. (1966) *Non-Standard Analysis*. Amsterdam: North Holland.
Salmon, W. C. (2001) Introduction. In Salmon, W. C. (ed.) *Zeno's Paradoxes*. Indianapolis, IN: Hackett Publishing Company.
Seidenfeld, T., Schervish, M. J. and Kadane, J. B. (2001) Improper Regular Conditional Distributions. *The Annals of Probability.* 29. 4. pp. 1612–24.
Shafer, G. (1976) *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.
Spielman, S. (1977) Physical Probability and Bayesian Statistics. *Synthese*. 36.2. pp. 235–69.
van Fraassen, B. (1989) *Laws and Symmetry*. Oxford: Oxford University Press.
Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman Hall.
Williamson, J. (1999) Countable additivity and subjective probability. *The British Journal for the Philosophy of Science.* 50. 3. pp. 401–16.
Williamson, T. (2007) How Probable is an Infinite Sequence of Heads? *Analysis.* 67. 295. pp. 173–80.

CHAPTER 9

CONDITIONAL PROBABILITY

KENNY EASWARAN

9.1 Introduction

As explained in previous chapters, a probability function is a function that is non-negative, (finitely or countably) additive, and normalized (that is, the most probable events have probability 1). There are many different interpretations of probability (see the section on interpretations, Chapters 15 to 22 in this volume), but they share much of the formalism, though there is some variability in whether the objects that bear probability values are taken to be sets of some sort, or sentences. I will generally use sentential notation in this chapter, except in sections 9.4.3 and 9.4.4, where the set-theoretic notation will be needed. At any rate, for neutrality, I will refer to the bearers of probability as "events", despite the many different interpretations they might have, and I will refer to the collection of all events as an "algebra", because of the logical or set-theoretic structure that it has. (In a few cases, it will need to be thought of as a "σ-algebra", which is an algebra that has not just negation and finite disjunctions and conjunctions, but also disjunctions and conjunctions of countably infinite collections of events.)

On any formulation of probability, it is common to take conditional probability to be defined in terms of unconditional probability. For any event B whose probability is non-zero, it is standard to define:

$$P(A|B) = P(A \wedge B)/P(B)$$

On a set-theoretic approach, "$\wedge$" is replaced by the set-theoretic intersection. I will call this the "ratio definition" of conditional probability.

Whenever $P(B) > 0$, one can define a function P_B, where $P_B(A) = P(A|B)$, for every event A. Under the definition given above, one can show that in such a case, P_B is itself a probability function, and that $P_B(B) = 1$. In some sense, one can think of P_B as the probability function that is "most similar" to P out of all the probability functions on the same algebra that assign 1 to B.

On a frequency interpretation of probability, $P(A|B)$ represents the fraction of things of type B that are also of type A. On a logical interpretation of probability, $P(A|B)$ is often

taken to represent something like the degree to which B entails A. On an interpretation of probability as degree of belief, the conditional probability function P_B is taken to represent the degrees of belief that an agent ought to have after learning B. On an evidential interpretation of probability, P_B is often taken to represent the probabilities where B is the total evidence.

Although this standard definition of conditional probability is good enough for many applications of probability, it is generally taken to be incomplete, and sometimes even to yield the wrong numerical values in certain cases. Notably, the standard definition is not applicable in cases where the event being conditioned on has zero, imprecise, or undefined probability.

In the rest of this chapter I will consider some of these issues that arise in the theory of conditional probability. In Section 9.2, I will consider how conditional probability relates to the indicative conditionals of natural language. In Section 9.3, I will consider some problems for conditional probability that are specific to probability interpreted as chance. And in Section 9.4, I will discuss some of the long and complicated history of alternative theories of probability that are intended to deal with the cases of probability 0, which motivate a treatment of conditional probability as the basic notion, rather than the defined one.

9.2 The Probability of Conditionals and Conditional Probability

Conditional probability obviously has some sort of connection to the conditionals of natural language. One thought is that the statement $P(B|A) = x$ means the same thing as $A \rightarrow (P(B) = x)$, where $\rightarrow$ is some ordinary conditional. That is, to say that the probability of B given A is x is just to say that if A is true, then the probability of B is x. However, some simple reasoning shows that this interpretation can't be right. Since $P(A|A) = 1$ and $P(A|\neg A) = 0$, then we would get a kind of fatalism—if A is true, then its probability is 1, and if it is false, then its probability is 0, and there is no room for intermediate probability.

For probability interpreted as degree of belief, a version of this view can be saved. One might say that $P(B|A)$ doesn't represent the value that $P(B)$ has if A is true, but rather the value that $P(B)$ *should* have if the relevant agent were to *learn* A (and nothing more). Because of a passage in Ramsey (1931), this condition is often known as the "Ramsey test". On this picture, the new unconditional probability represents the (possibly counterfactual) degree of belief of the agent in a situation with more evidence than the original conditional probability. Thus, this is an explication just as much of the relation between the two different probability functions as of the notion of conditional probability.

Another suggestion that has been made is that the conditional probability $P(B|A)$ is equal to the probability of the conditional $P(A \rightarrow B)$ (Adams, 1965, Stalnaker, 1970, Adams, 1975, 1996). However, to understand this equation, we need to understand the conditional $\rightarrow$.

It is clear that if $\rightarrow$ represents the material conditional from classical logic, then this equation simply fails to hold in most cases. On this interpretation, $P(A \rightarrow B) = P(A \wedge B) + P(\neg A)$, which is $P(B|A) \cdot P(A) + 1 \cdot P(\neg A)$. Since $P(B|A) + P(\neg B|A) = 1$, this means that $P(A \rightarrow B) = P(B|A) \cdot P(A) + P(B|A) \cdot P(\neg A) + P(\neg B|A) \cdot P(\neg A)$. Since $P(A) + P(\neg A) = 1$,

this means that $P(A \rightarrow B) = P(B|A) + P(\neg B|A) \cdot P(\neg A)$. Thus, this can only equal $P(B|A)$ if $P(A) = 1$ or if $P(B|A) = 1$.

Thus, since many authors have found the equation plausible, much research been has focused on finding an interpretation of the conditional on which it can be true. A series of mathematical results (Hájek and Hall (1994), Hájek (2012), and elsewhere), starting with Lewis (1976), have shown further triviality results of this sort. Various relatively weak assumptions about the logic of conditionals, together with the equation of conditional probability with the probability of conditionals, entail that the probability function is restricted to only four values. Thus, to hold on to the equation, one must give up many plausible principles in the logic of conditionals, possibly including the very idea that sentences involving conditionals can be either true or false (Edgington, 1995)! Thus, defenders of the equation are more likely to use conditional probability to illuminate conditionals than the other way around.

Others have just accepted that the connection between conditional probability and the conditionals of natural language is not as tight as one might have thought. Conditional probability tells us something about how the agent will update in the light of new evidence, while conditionals from natural language tell us something different.

9.3 Humphreys' Paradox

Paul Humphreys raised a special problem for conditional probability on the chance interpretation (Humphreys, 1985). Given the ratio formula for conditional probability, if $P(A)$ and $P(B)$ are both positive, then

$$P(A|B) = P(B|A) \cdot P(A)/P(B).$$

This formula is known as "Bayes' Theorem", and has been considered quite important in several interpretations of probability. In the case of chance, it means that conditional probabilities can be "inverted"—if the probabilities of two events A and B are well-defined and positive, and $P(B|A)$ is as well, then so is $P(A|B)$. However, for many applications of chance, this seems to conflict with an intuitive understanding of what conditional probability means.

It is natural to think that on the chance interpretation, $P(A|B)$ tells us something like the degree to which a situation of type B tends to produce a situation of type A. This is bolstered by the observation that specifications of chances seem to always rely on a specification of background conditions. The chance of a coin landing heads requires a specification of the conditions under which the coin was flipped, to indicate whether the flip or the coin had any bias. It is natural to suppose that this specification of the conditions requires a conditional probability. Thus, in the formula "$P(A|B)$", we can think of "B" as specifying the conditions of the chance setup, and "A" as specifying the event whose chance in such a situation we are interested in.

However, if we interpret the formula in this way, then it seems that $P(A|B)$ and $P(B|A)$ can't both make sense. Either B is the specification of conditions under which events of type A might occur, or vice versa, but there seem to be very few situations in which both could be

the case. One can talk about the chance of a flip coming up heads, given the specifications of the conditions under which the coin was flipped, but the fact that the coin comes up heads doesn't seem sufficient to specify the chance of the coin having been flipped in a particular way. This motivates the thought that at least for the notion of chance, a very different formalism for conditional probability is required.

Christopher McCurdy suggests that this argument rests on a mistake about the notion of conditional probability involved (McCurdy, 1996). He argues that the background conditions should not be included as the conditioning event in a standard conditional probability formula, but should instead be taken as part of what defines the probability function itself. That is, if B is the specification of background conditions, then the chance of A under such conditions should be written as something like "$P^B(A)$" rather than "$P(A|B)$". The conditional probabilities that do make sense instead represent information about the correlations among events that both have some tendency or other to be produced by a single experimental setup. For example, in the coin flip case, one might ask about the chance of both flips coming up heads, given that at least one flip does, in a case where two fair coins are about to be flipped independently. This would be represented as something like $P^B(A|C)$, where B is the condition that two fair coins are about to be flipped independently, A is the condition that both land heads, and C is the condition that at least one flip does. McCurdy then argues that these conditional probabilities can be understood with the standard ratio formula, so that the problem is resolved.

But this resolution raises further questions. Conditional probability may always be given by the ratio definition, but the question now arises of how chances given one background relate to chances given another background. One might wonder what relations hold between the chance of B, and the chance, in a situation set up as type B, of A. That is, just as the ratio formula gives some calculations that allow us to shift certain events back and forth across the vertical slash of traditional conditionalization, a question remains about whether there is any calculation that allows us to shift certain events from the superscript describing the background conditions of one probability function into the domain of another probability function. I take it that this is the serious challenge raised by Humphreys (2004) to the solution proposed by McCurdy. What is the appropriate formalism relating probability on one set of background information to probability on another set of background information? Some approaches are given in the work of Isaac Levi (for instance, see his (1977)), but a full theory has yet to be worked out.

9.4 Conditional Probability as Basic

Similar concerns about background information have been raised for other interpretations of probability. For instance, in the concluding section of his 2003, Alan Hájek says about one's credence of 1/2 in a coin landing heads: "really it is conditional on tacit assumptions about the coin, the toss, the immediate environment, and so on. In fact, it is conditional on your total evidence" (Hájek, 2003, p. 315). And he cites a similar claim from de Finetti: "every evaluation of probability, is conditional; not only on the mentality or psychology of the individual involved, at the time in question, but also, and especially, on the state of information in which he finds himself at that moment" (de Finetti, 1974, p. 134).

However, I claim that this is a confusion analogous to the one that leads to Humphreys' paradox. There is a distinction between the credence that one actually has, which happens to take into account all of one's evidence and biases, both explicit and implicit, and the credence one has when explicitly making some further supposition. The latter is what conditional probability is traditionally taken to model. The former is again some notion of probability on a "background" of some sort.

Subjective Bayesians should be especially firmly committed to this point. On the subjectivist view, two rational agents may have different credences, despite having exactly the same evidence. The difference between these credences must be represented by using a different probability function, rather than putting a further proposition into the conditional probability. If Jane and John have different credences, despite having the same evidence, one wouldn't want to say that these credences are $P(A|Jane)$ and $P(A|John)$ (what would the unconditional P even mean?!), though one might say they are $P^{Jane}(A)$ and $P^{John}(A)$. To write it as the former is to somehow suggest that being Jane or being John has some evidential role, and that Jane and John somehow recognize each other as having some reason for having different credences.

In cases where the difference between agents also involves a difference in background beliefs, which may consist in biases, unstated assumptions, presuppositions, or other attitudes of which the agent may not even be aware, similar reasoning applies. If John has a background belief B about how the world works that he has never articulated, one wouldn't necessarily want to say that his credence in A is $P(A|B)$—that would suggest that if one were to articulate B and tell it to someone who shared all the rest of John's background assumptions, then this person would come to have the same credence in A. But this seems implausible—even John himself might change his views on becoming consciously aware of these background beliefs B, and so there is no reason to expect someone else to come to have his views on learning these beliefs. B may play a role in generating John's credence in A, but that doesn't mean that his credence in A is actually a credence conditional on B. And it doesn't mean that two people with the same background assumptions will have the same credences—there may be further non-propositional features of the mental states that make them differ.

Thus, these sorts of considerations at best motivate the claim that all probability involves some sort of appeal to a background. However, there are other concerns that suggest that we need some way to make sense of conditional probabilities that doesn't depend on deriving them from unconditional probabilities.

9.4.1 Undefined Ratios

Although Hájek uses the previous sort of argument at the end of his (2003), the bulk of this paper is devoted to arguing that there are many clear examples of propositions whose probability is either zero, imprecise, or undefined, for which we can nevertheless make sense of precise non-trivial conditional probabilities involving these propositions.[1] Since the traditional ratio definition of conditional probability says that $P(A|B) = P(A \wedge B)/P(B)$,

[1] There is also a discussion of "infinitesimal" probabilities, which involve non-standard number systems on which there are numbers that are larger than zero but smaller than any positive real number.

this means that $P(B)$ must be well-defined, precise, and positive, in order for the conditional probability to be defined. Thus, these cases appear to be counterexamples to the traditional ratio definition. Hájek argues that cases of all of these sorts can arise both for chance and for credence, so that for both interpretations (and presumably many other related ones as well), conditional probability must be analyzed in some way other than by the ratio definition. He suggests that an appropriate analysis will eventually take conditional probability to be basic, rather than defined in terms of unconditional probability.

As a typical example of the probability zero problem, consider a situation in which an agent is about to throw a dart at a dartboard. She is wondering about the probability that the center of the dart will land precisely on the vertical line bisecting the board. Given that the agent doesn't have infinite precision in her throwing abilities, the precise center of the dart could end up on infinitely many vertical lines on the board. For any strip of finite width including the center strip, there will be some positive probability that the dart hits that strip, but one can break the strip into several smaller ones that are approximately equally likely, and each of those can be split further. Thus, we can find strips of positive width that have arbitrarily low probability, and therefore, the probability of hitting precisely the infinitely thin strip that bisects the board must be less than any of them, and therefore 0.

And yet, we can make sense of the conditional probability that the dart hits the upper half of the board, given that it is precisely centered. If the dart throw is relatively unbiased, then in fact it seems plausible that this conditional probability should be $1/2$. And yet, the traditional definition (if it tells us anything) will tell us that the probability that the dart hits the upper half, given that it hits the center line, is the probability that it does both (which is 0), divided by the probability that it hits the center line (which is also 0). And yet $0/0$ is mathematically undefined, and so the traditional formula fails to give the right (or indeed, any) answer. In the three subsections to come, I will discuss mathematical theories of conditional probability that allow for these conditional probabilities to be defined. All of these theories accept that when $P(B) > 0$, then $P(A|B) = P(A \wedge B)/P(B)$, but they additionally either allow or even require that $P(A|B)$ be defined when $P(B) = 0$. On two of them (those due to Popper and to Rényi), conditional probability is in fact the basic notion, as Hájek suggests, but on the third (due to Kolmogorov) it isn't—it is still defined in terms of unconditional probability.

Hájek raises further worries about cases where some of the relevant unconditional probabilities are imprecise (he uses the term "vague") or undefined. (For more on imprecise probability, see Chapter 14 by Fabio Cozman in this volume.) Hájek argues that on various applications of probability, there may be events whose conditional probability is well-defined, even though their unconditional probabilities are not. (Imprecise or vague probabilities come on pp. 293–15, and undefined probabilities on pp. 295–312.)

The approach to imprecise probability that he discusses is one on which impreciseness is represented "supervaluationally"—instead of one precise probability function that represents the relevant application of probability, there is a set of such functions that together represent the probability. The facts about the probability are just the facts that are shared by all of these functions, while facts that differ between the functions are in some sense

However, this case poses less of a direct problem for defining conditional probability in terms of unconditional, and is beyond the scope of the present chapter. See further discussion in Easwaran (2014).

indeterminate. In the cases Hájek gives, it is natural for every probability function in the set to give the same conditional probability (at least, once the theory of conditional probability is modified to deal with probability 0), so that it is precise, even though the unconditional probabilities are different, and thus imprecise. Although this supervaluationism is not itself a component of the various formalisms for conditional probability that I will discuss, it can be straightforwardly added to any of them without any new problems.

I will not rehearse the arguments involving undefined probabilities here, but will instead direct the interested reader to Hájek's paper. Some examples rely on the controversial idea that an agent can't have a credence in her own free action, and yet she can have credences conditional on (for instance) her free choice of whether or not to flip a coin. Other examples rely on the idea that even for a "non-measurable" set A, the agent has a conditional credence in A given A, which is 1. (One might say that in these cases, whatever prevents the agent from having an unconditional credence also prevents her from having a conditional credence, even though the value that she would have, were she able to have such a conditional credence, is obviously 1.) If the arguments are good, then they seem to vindicate Hájek's view that conditional probability must be basic, rather than being defined in terms of unconditional probability. However, they also cause problems for all three formalisms I will discuss. Popper and Kolmogorov both require that all unconditional probabilities are well-defined, while Rényi requires either that all are or that none are. Thus, if the problem of undefined probabilities is important, then it calls for new mathematics. But for now, I will discuss these three prominent formalisms for dealing with the problem of zero probability, all of which should be able to be easily "supervaluated" to account for imprecise probability.

9.4.2 Popper

Perhaps the solution to the problem of probability 0 that philosophers most commonly propose is the axiomatization by Karl Popper (Popper, 1955, 1959). Popper assumes only that the objects of the probability function have a binary and a unary operation defined on them (which can later be interpreted as conjunction and negation), that for any two such objects their conditional probabilities both exist and are real numbers, and that the values of the conditional probability function satisfy the following set of axioms:

A1 $P(X \wedge Y|Z) \geq P(Y \wedge X|Z)$
B1 $P(X|Z) \geq P(X \wedge Y|Z)$
B2 $P(\neg Z|Z) + P(X|Z) = P(X \wedge Y|Z) + P(X \wedge \neg Y|Z)$
B3 $P(X \wedge Y|Z) = P(X|Y \wedge Z)P(Y|Z)$
C1 $P(X|X \wedge Y) = P(Y|Z \wedge Y)$
C2 If $P(\neg Z|Z) \neq 0$ then $P(X|Y) = P(X|Y \wedge \neg Z)$
D1 If $P(\neg Y|Y) \neq 0$ then $P(X) = P(X|\neg Y)$
E1 There are X and Y with $P(\neg X|X) \neq P(\neg Y|Y)$

(This particular formulation is taken from Popper (1955), which has a typo omitting a negation sign in D1. A somewhat different formulation is presented and discussed fairly extensively by James Hawthorne in Chapter 13 of this volume.)

This system has some very nice advantages over the traditional Kolmogorov axiomatization. One is that nothing is assumed about the entities that bear the probabilities, so one can take them to be sentences, sets, concrete historical events, or anything else, as long as one can interpret the two operations appropriately. Additionally, the meaning of the two operations is not directly assumed, but is in some sense derivable from these axioms—that is, if two expressions are tautologically equivalent under the standard interpretation of "$\wedge$" and "$\neg$", then one can show merely using these axioms that substituting one expression for the other in some formula of probability will give the same value. Additionally, one can define a sort of probabilistic entailment relation, where X "P-entails" Y iff for all Z, $P(X|Z) \leq P(Y|Z)$, and one can show that P-entailment is an extension of the classical tautological entailment relation for any Popper function P.

Perhaps most importantly, these axioms entail the standard probability axioms, and when $P(Y) \neq 0$, they entail the ratio formula, that $P(X|Y) = P(X \wedge Y)/P(Y)$. However, $P(X|Y)$ still always has a value, even if $P(Y) = 0$. For some Y, often called "normal", one can show that the function P_Y (where the unconditional function is defined, as earlier, by $P_Y(X) = P(X|Y)$, while the conditional function is defined by $P_Y(X|Z) = P(X|Y \wedge Z)$) is itself a Popper function. For any Y that is not normal (often called "abnormal"), one can show that $P(X|Y) = 1$ for all X—in particular, this means that $P(Y|Y) = P(\neg Y|Y) = 1$. The conditions in axioms C2 and D1 are there to limit their application to abnormal elements, and E1 guarantees that there are both normal and abnormal elements. These abnormal elements play the role of contradictions, and one can show that any syntactically contradictory element is abnormal, but there may be other abnormal elements as well. One can show that every abnormal element has probability 0, but there may be normal elements of probability 0 as well, and this possibility is what allows the system to get around the problem of probability 0.

These Popper functions have been compared to a few other methods for defining conditional probability in these probability 0 cases, namely probability functions that use infinitesimals, and "lexicographic probabilities" (van Fraassen (1995), McGee (1994), Halpern (2010)). For every Popper function, there is a function of each of the two other sorts that represents it in some relevant sense. However, since the representation is not always unique, there is some difference among these different structures.

For Hájek's purposes however, Popper functions will not work. First of all, they require that every event have both conditional and unconditional probabilities, and thus they can't accommodate the cases of undefined probability that he is interested in. Additionally, one of Hájek's proposed cases with probability 0 causes problems for Popper functions. Hájek wants to be able to conditionalize on events of probability 0 in order to make sense of probabilities conditional on things that the agent has already learned to be false.

> The die is tossed, and you conditionalize on the evidence that the die landed odd. You then assign probability 0 to the die landing with an even face showing up on that toss. Still, surely rationality permits you to assign P(die landed even, given die landed even) = 1 (and, I would add, even requires it). (Hájek, 2003, p. 288)

Let E be the event that the die lands even, and $\neg E$ be the event that the die lands odd. Let Q be one's credence function before observing how the die landed, and P be one's credence function after conditionalizing on the evidence that the die landed odd. The natural generalization of conditionalization says that $P = Q_{\neg E}$, so that for every X and Y,

we should have $P(X|Y) = Q(X|Y \wedge \neg E)$. Thus, we should have $P(E|E) = Q(E|E \wedge \neg E)$, and since $E \wedge \neg E$ entails E, this means that $P(E|E) = 1$, as Hájek wants.

However, there is a problem. If we let X be any arbitrary event unrelated to the die roll, then we have $P(X|E) = Q(X|E \wedge \neg E)$, and since $E \wedge \neg E$ is a syntactic contradiction (and thus, abnormal), we still have $P(X|E) = 1$. In particular, we get $P(\neg E|E) = 1$! I suspect that Hájek would want it to be the case that $P(\neg E|E) = 0$.

At any rate, this certainly won't do for two other philosophers who are interested in similar cases of probability 0. Joyce (1999, pp. 200–15) and Christensen (1999, p. 450) both want to solve the problem of old evidence for confirmation theory by including probabilities conditional on the negation of one's old evidence. (For more on this problem, see Chapter 27 on Confirmation Theory by Vincenzo Crupi and Katya Tentori, in this volume.) In particular, they point out that if E is some evidence that one has already learned, then traditional measures of confirmation say that the degree to which E confirms any hypothesis whatsoever is 0. They propose that we replace the traditional measures by a new measure $S^*(E,H) = P(H|E) - P(H|\neg E)$, and use Popper functions to make sense of the conditional probabilities even in cases where E has already been learned as evidence.

Interestingly, although they show that this measure gives reasonable results when $P(E)$ is extremely close to 1, they don't work through the details of what happens when E has already been learned as evidence. As in the previous paragraph, if E has been learned, then $\neg E$ will be abnormal, and so $P(H|\neg E) = 1$. Thus, on this measure, if E is evidence that has been learned by traditional conditionalization, then $S^*(E,H)$ is never positive—old evidence can only ever disconfirm by this measure (and in fact, it counts as disconfirming both H and $\neg H$, if $0 < P(H) < 1$).

Both Joyce and Christensen argue that there is no one measure of confirmation that properly deals with all cases, so that their preferred measures may go wrong for some purposes, and should be replaced by other measures in those cases. However, for the problem of old evidence, Popper functions won't help. Either one doesn't need to appeal to probability 0 (in which case one can use the traditional axioms), or one is conditionalizing on the negation of the old evidence (which is abnormal and thus always gives 1). Similarly, they will be of no help in considering some sort of "counterfactual" conditional probability (which may play a role in causal decision theory, as endorsed by Joyce (1999)), where one imagines what the probabilities would be, were something false that one currently knows to be true. For these cases, we may need something more like the view proposed (but not fully worked out) in Levi (1977).

9.4.3 Rényi

Another proposal that is often mentioned in the same breath as Popper's is the probability theory due to Alfréd Rényi (Rényi, 1955). However, the motivating cases for Rényi's theory are quite different from the ones for Popper's theory, and though the two are referred to similarly, Rényi's theory is much more general. This proposal involves more complicated mathematical techniques from real analysis and set theory, partly because it was developed to deal with some situations that arise in mathematical statistics. As a result of these complications, many philosophers have failed to notice that it doesn't address the problems mentioned above any better than Popper's theory.

Rényi's system starts with a conditional probability function P that is defined on two collections of subsets of a set Ω. As in the standard set-theoretic approach, $\mathcal{A}$ is a σ-algebra of subsets of Ω, but there is a distinguished subset $\mathcal{B} \subseteq \mathcal{A}$. Rényi requires that $P(A|B)$ is defined iff $A \in \mathcal{A}$ and $B \in \mathcal{B}$. The function additionally must satisfy the following three conditions:

1. $P(A|B) \geq 0$ whenever $A \in \mathcal{A}$ and $B \in \mathcal{B}$, and $P(B|B) = 1$ whenever $B \in \mathcal{B}$.
2. For any fixed $B \in \mathcal{B}$, $P(\ldots|B)$ is countably additive. (That is, for any $A_1, A_2, \ldots$ in $\mathcal{A}$, $P(\bigcup A_i|B) = \sum P(A_i|B)$.)
3. If $A, B, C \in \mathcal{A}$ and $C, B \cap C \in \mathcal{B}$, then

$$P(A|B \cap C) \cdot P(B|C) = P(A \cap B|C).$$

From these axioms, it is straightforward to prove that for any $B \in \mathcal{B}$, $P(\ldots|B)$ is in fact a traditional probability function, and that for any traditional probability function, the ratio definition gives rise to a Rényi function where $\mathcal{B}$ is taken to be the set of all sets with non-zero probability. Additionally, any countably additive Popper function on a σ-algebra of sets gives rise to a Rényi function, where B is the set of all normal events.

However, Rényi's interest in these functions was prompted by a different class of cases where B doesn't just exclude sets for being too small (probability 0, or abnormal), but also excludes some sets for being too large. As a standard example of such a space, consider the algebra $\mathcal{A}$ of all Lebesgue-measurable subsets of the set $\mathbb{R}$ of all real numbers. Let μ be the standard Lebesgue measure on this collection of sets. (For those that are unfamiliar with the notion, Lebesgue measure is just the straightforward measure that one gets by letting the measure of an open or closed interval be the difference between the two endpoints, and extending this in the natural way to countable unions and intersections.) Let $\mathcal{B}$ be the set of Lebesgue-measurable sets of real numbers that have positive, but finite, Lebesgue measure. Then, let $P(A|B) = \mu(A \cap B)/\mu(B)$. One can check that this function satisfies Rényi's axioms.

Unlike Popper functions, this example doesn't have a natural notion to represent unconditional probability. For a Popper function, the unconditional probability is defined as the probability conditional on the negation of an abnormal event. But in this function, probability conditional on the empty set and on the whole space are both undefined, since $\mu(\emptyset) = 0$ and $\mu(\mathbb{R}) = \infty$. Thus, neither of these complementary sets is in $\mathcal{B}$.

One can try to define an unconditional probability $P(A)$ as the limit of $P(A|B)$ as B gets larger and larger, approaching $\mathbb{R}$. However, this will end up telling us that whenever $\mu(A)$ is finite, $P(A) = 0$. Since these are the only events that we can conditionalize on, we see that we end up having said that the *only* events one can conditionalize on are ones with probability 0, rather than merely allowing them as special cases. This extension also forces us to give up on countable additivity (since there is a countable collection of unit intervals that together cover $\mathbb{R}$). If we try to force in probabilities conditional on other events that aren't in $\mathcal{B}$, these problems become even more serious. Thus, it is much more natural to treat this case in Rényi's way, where one only considers conditional probabilities, and only ones with particular conditioning events, namely the ones in $\mathcal{B}$.

It turns out that examples like this one arise naturally in many statistical problems, where they are often called "improper distributions". The second half of Rényi's long paper shows how they come up in various applications, such as summing infinitely many independent

random variables, calculating Maxwell's distribution of the velocities of molecules in an infinite gas, and various other applications where we are interested in local probabilities of events in unboundedly large spaces. It is precisely for these examples that Rényi gave his theory of conditional probability. There don't appear to be any ways to use this theory to allow for conditional probabilities on events that don't have unconditional probabilities, except in this sort of case, where *no* events have non-trivial unconditional probabilities.

Rényi's treatment of conditional probability as a basic notion is somewhat more general than Popper's. If we treat Popper's constant probability 1 conditional on abnormal events as being merely a notational variant of a lack of conditional probabilities, then Popper functions are a special case of Rényi functions. However, the generalization is of no help for the philosophical problems that Popper functions failed to solve. Abnormal events are more stubborn, and there is no way for a conditional probability function to take some inputs that lack an unconditional probability, unless all of them do.

9.4.4 Kolmogorov

Although Popper and Rényi both allow for probabilities conditional on events whose unconditional probability is 0 (and Popper even requires that such probabilities exist), neither formalism gives any constraint on how those particular conditional probabilities relate to unconditional probabilities. Both formalisms include a multiplicative version of the ratio definition, saying that $P(A \wedge B) = P(A|B)P(B)$, and when $P(B) = 0$, we already know that $P(A \wedge B) = 0$, so this equation tells us nothing about $P(A|B)$. However, there are older mathematical results that give an approach to calculating such conditional probabilities, which in a sense allow us to continue to take unconditional probability as the basic notion.

The basic idea is a generalization of some observations that we can make in the case of conditionalizing on events of positive probability. The Law of Total Probability (which is just a consequence of the ratio definition of conditional probability, and finite additivity) says that if events $E_1, \ldots, E_n$ form a partition (that is, they are mutually exclusive and exhaustive, so that it is guaranteed to be the case that exactly one of the E_i is true), then

$$P(A) = P(A|E_1)P(E_1) + \cdots + P(A|E_n)P(E_n).$$

Since the $P(E_i)$ add up to 1, this means that $P(A)$ is a weighted average of the $P(A|E_i)$. This has two interesting consequences, which seem to be plausible candidates for requirements on any appropriate extension of conditional probability.

The first consequence is known as "conglomerability". It says that

$$\min P(A|E_i) \leq P(A) \leq \max P(A|E_i).$$

It turns out that neither Popper's nor Rényi's axioms entail this condition for infinite partitions, even given the appropriate generalization of min to inf and max to sup.[2] To see that this can fail on Popper's axioms, consider an infinite partition where each element

[2] An infinite set may not have a highest or lowest value, so it is natural to instead consider, instead of the minimum value, the greatest number such that every element of the set is higher than it, which is the

has unconditional probability 0 (for instance, the space may be a square dartboard, and the partition may be into the set of infinitely thin vertical lines), and let every event with unconditional probability 0 be abnormal. In this case, if A is any event such that $P(A) < 1$, then there is a violation of conglomerability for this partition, since the probability of A conditional on any abnormal event is 1, so all conditional probabilities equal 1, and yet the unconditional probability is lower. Examples of a violation in Rényi's system are less straightforward, but they are implied by some of our later results.

The second consequence is known as "disintegrability". To state this consequence, I will define a random variable $V^{A,\mathcal{E}}$, where $\mathcal{E}$ is any partition, and A is any event.[3] Let $V^{A,\mathcal{E}}(s)$ be the value of $P(A|E_i)$ for the unique element E_i of $\mathcal{E}$ that contains s. The probability function P is said to be disintegrable in a partition $\mathcal{E}$ iff for every event A, $P(A)$ is the expectation of $V^{A,\mathcal{E}}$. This generalizes the observation for finite partitions that the unconditional probability is a weighted average of the conditional probabilities, since "expectation" is the natural generalization of weighted average. Since the expectation of any random variable is always between the inf and sup of the values of the random variable, disintegrability entails conglomerability. And it turns out that the converse is true, so that the two conditions are equivalent (Dubins, 1975).

And it turns out that, although many philosophers consider Kolmogorov to have *defined* conditional probability in terms of a ratio, in chapter 5 of his (1933/1950), he considered something very much like the Popper and Rényi generalizations, together with the assumption that conditional probability be disintegrable in various partitions. For any finite partition, it is straightforward to check that conglomerability and disintegrability are equivalent to the standard ratio definition, for all events whose unconditional probability is positive. But to generalize to events whose unconditional probability is 0, Kolmogorov used the Radon-Nikodym theorem to show that for any countably additive unconditional probability function, and any partition, there is a set of conditional probabilities on that partition that satisfy conglomerability and disintegrability.

However, it turns out that imposing this requirement puts some further constraints on conditional probability, which make it in some ways substantially different from anything compatible with either the Popper or Rényi axioms.

The first condition that is required for conglomerability in a given partition is that the probability function be countably additive in this partition. It is proved in both Schervish et al. (1984) and Hill and Lane (1985) that if a probability function on a countable space fails to be countably additive, then there is no conditional probability function extending it that satisfies conglomerability. As a typical example, consider a probability function whose space is the set of natural numbers, such that $P(A|B)$ is defined as the limit, as n goes to infinity, of the relative frequency of members of set A among the first n members of B. Let E_i be the set of numbers that are powers of i, where i is any natural number that is not itself the square, cube, etc. of any other natural number. These E_i form a partition of the natural numbers. Let S be the set of perfect squares. Now for any E_i, $P(S|E_i) = 1/2$, because every

"inf", and instead of the maximum value, the lowest number such that every element of the set is lower than it, which is the "sup".

[3] A random variable is some quantity that depends on which state is actual. It is most straightforward to define it on the set-theoretic formulation of probability as a function $V\colon \Omega \to \mathbb{R}$, such that each set $V_x = \{s|V(s) > x\}$ is itself an event.

other power of i is a perfect square. And yet $P(S) = 0$, since the perfect squares get less and less common among the larger numbers. It is not obvious from considering this case that similar problems must occur for any probability function that fails to satisfy countable additivity, but it turns out to be true.

A further problem is already remarked upon by Kolmogorov, who refers to it by the traditional name of the "Borel paradox", though it was actually first introduced in Bertrand (1889). Consider the uniform probability distribution over the surface of a sphere, where the probability of any ordinary region is proportional to the area of that region. Fix one axis of the sphere, and call the endpoints the "north pole" and "south pole", and consider the partition of the sphere into the set of great circles through these points, which we can call the "lines of longitude".[4] Similarly, we can consider the circles that are everywhere perpendicular to the lines of longitude, and number them as "lines of latitude", exactly as we do on Earth. Let A be the region of the sphere consisting of the points with latitude between θ_1 and θ_2, with the equator counting as 0. One can show that $P(A) = |\sin\theta_1 - \sin\theta_2|/2$.

Now, we can consider the question of what $P(A|E)$ is, for a given line E of longitude. On the one hand, the symmetry of the problem suggests $P(A|E_1) = P(A|E_2)$, where E_1 and E_2 are any two lines of longitude. This means that for the partition into the lines of longitude, every conditional probability of A must be the same. But then, in order for conglomerability to hold, these conditional probabilities must equal the unconditional probability, so that $P(A|E) = |\sin\theta_1 - \sin\theta_2|/2$. If A' is any region that is not itself composed of lines of latitude, but that intersects E in exactly the same points as A does, the conditional probability function must give the same result. (Every theory of probability we have considered requires that $P(A|E) = P(A \cap E|E)$, and in this case $A \cap E = A' \cap E$.)

As Kolmogorov notes, this means that the unconditional *uniform* probability distribution on the surface of the sphere gives rise to a *non*-uniform probability distribution conditional on each line of longitude. (The uniform conditional distribution would say that $P(A|E) = |\theta_1 - \theta_1|/\pi$.) Furthermore, the initial choice of axis was arbitrary, and each line of longitude for one axis is also a line of longitude when any other point on the circle is chosen as one of the poles. But then satisfying symmetry and conglomerability around a different axis will give a *different* non-uniform probability distribution conditional on the same line of longitude. This is the paradox.

Some philosophers and statisticians (particularly those influenced by de Finetti, and his opposition to countable additivity as a requirement on probability) respond by giving up conglomerability as a requirement on conditional probability. However, to me it seems more fruitful to maintain conglomerability, and reconceptualize the notion of conditional probability. Rather than considering conditional probability to be a fixed two-place function $P(A|B)$, conditional probability can be relativized to the partition—if $\mathcal{E}_1$ and $\mathcal{E}_2$ are two

[4] Strictly speaking, this collection of great circles is not a partition, since the lines all overlap at the poles. Because this overlap is so small compared to the individual sets in the partitions, it turns out not to matter. By removing the poles, and a small set of other points, this example can be fixed so that the approximate partitions in the example are in fact precise partitions.

A different example that illustrates the same point, and avoids overlap among elements of the "partitions" is described in Example 6.1 of Kadane et al. (1986). The example there involves two independent normally distributed random variables X and Y, and the partition of the sample space into sets with constant Y/X, which is in some sense no more complicated than a uniform distribution over a sphere and its "partition" into lines of longitude.

different partitions, and B is an element of both partitions (for instance, if B is a line of longitude, and $\mathcal{E}_1$ and $\mathcal{E}_2$ are the partitions into the lines of longitude given by two choices of north pole, both of which are on B), then $P^{\mathcal{E}_1}(A|B)$ may not equal $P^{\mathcal{E}_2}(A|B)$. This idea is suggested in chapter 5 of Kolmogorov (1933/1950), and more directly in the contributions by Dickey, Fraser, and Lindley in the addenda to Hill (1980). This is a modification of the framework that is quite different in character from the suggestions of Popper and Rényi.

This modification is quite significant for cases like the Borel paradox, because it says that there is no such thing as the value $P(A|B)$, when B is a line of longitude—instead, there is only $P^{\mathcal{E}}(A|B)$, for various partitions $\mathcal{E}$. However, in cases where $P(B) > 0$, the modification is less significant. As mentioned above, conglomerability requires that in such a case, $P^{\mathcal{E}}(A|B) = P(A \wedge B)/P(B)$, regardless of which partition $\mathcal{E}$ one is using. Thus, one can drop the relativization to the partition in cases where one conditions on an event of positive probability. The relativization affects only probabilities conditional on events of probability 0. It is an extension of the ratio definition to cases of probability 0, which introduces dependence on a new parameter for those cases, and not a contradiction of the ratio definition.

Arntzenius et al. (2004) propose one other potential problem for conglomerability. (There are several other problem cases mentioned in that paper, but they all involve non-measurable sets that don't have a probability, and sets without probabilities pose a more general challenge for the idea of conditional probability, which none of the present proposals address.)

For this case, consider a uniform probability over the unit square. Now let L be the left-most strip of width ϵ, and consider the set of vertical lines contained in L. Let R be the remainder of the square, and consider the set of points in this region. The cardinalities of the set of vertical lines in L and the set of points in R are the same, and thus we can pair them up, with one line and one point in each pair. These pairs then form a partition of the entire board. Imposing conglomerability over this partition gives some surprising requirements. In particular, since the unconditional probability that the point is in L is ϵ, there must be some pair such that the probability that the point is in L, given that it is in the pair, is at most ϵ. Yet Arntzenius et al. report the intuition that, since the line is in L and only one point is not in L, the probability that the point is in L given that it is somewhere on the pair should be 1, or very close to it.[5]

This is indeed a serious challenge to conglomerability, and I don't see any choice for the defender of conglomerability but to bite the bullet. Some such pair (and in fact, almost all of them) must have conditional probability less than ϵ, even though it seems that 1 is a more

[5] A similar example, involving finite additivity, is given on pp. 45–7 of Hill (1980). One knows initially that some physical constant m takes a rational value between 0 and 1, and splits one's credence between two hypotheses T_1 and T_2. Conditional on T_1, one's credences about m have a countably-additive distribution on which each rational number in the interval has positive credence. Conditional on T_2, one's credences about m are given by a uniform, finitely additive distribution over this interval, which assigns each particular value probability 0. The probability that m takes a particular value is thus positive, but the conjunction of this proposition with T_2 has probability 0, and thus conditional on the particular value of m, T_1 will have probability 1, which violates conglomerability.

Furthermore, Seidenfeld et al. (2014) show that a similar issue will arise for any continuous probability space, where some collection of events that individually have probability 0 jointly constitute an event of non-zero probability.

intuitive value. The only thing to ease the pain is the fact that these conditional probabilities of ϵ are only required relative to this particular partition—for other more natural partitions of the space, it is likely that the conditional probabilities will take more intuitive values. We only get these unintuitive values because we took an unintuitive partition of the space, using its cardinality features and not its geometric ones.

Kolmogorov's suggestion allows us to maintain conglomerability by making conditional probability depend on the partition, in addition to the two events involved. It also allows many conditional probabilities to be calculated from the corresponding unconditional probabilities. (In the case of the Borel paradox, this appealed to the symmetry of the sphere. In cases without such symmetry, the calculations may not be possible.) If this can be generalized, then it gives a way to deal with probability 0 without taking conditional probability to be the basic notion. However, it still calls for further investigation into the mathematical relations involving changes of partition, just as with modifications of "backgrounds" at the end of section 9.3.

9.5 Conclusion

In this chapter I have not gone into much detail on the applications of conditional probability. I have instead focused on the variety of formal theories of conditional probability that exist. The traditional ratio definition is good enough for cases where all relevant propositions have well-defined, positive unconditional probabilities. There are concerns about how to relate it to natural language conditionals, and to the role that a "background" plays in setting up chance or degree of belief. There are also more serious problems that arise when some events have probabilities that are zero, imprecise, or undefined. There are three formalisms available for dealing with probability 0, and all can be extended supervaluationally to imprecise probabilities. But if impreciseness must be treated in some way other than the supervaluational one, or if some probabilities can be undefined while the conditional probabilities are defined, then some further formalism will be required.

References

Adams, E. (1965) The logic of conditionals. *Inquiry*. 8. pp. 166–97.

Adams, E. (1975) *The Logic of Conditionals: an Application of Probability to Deductive Logic*. Dordrecht: Reidel.

Adams, E. (1996) *A Primer of Probability Logic*. Chicago, IL: CSLI.

Arntzenius, F., Elga, A., and Hawthorne, J. (2004) Bayesianism, infinite decisions, and binding. *Mind*. 113. 450. pp. 251–83.

Bertrand, J. (1889) *Calcul des probabilités*. Paris: Gauthier-Villars.

Christensen, D. (1999) Measuring confirmation. *The Journal of Philosophy*. 96. 9. pp. 437–61.

de Finetti, B. (1974) *Theory of Probability*, vol. 1. New York, NY: Wiley.

Dubins, L. (1975) Finitely additive conditional probabilities, conglomerability, and disintegrations. *The Annals of Probability*. 3. 1. pp. 89–99.

Easwaran, K. (2014) Regularity and hyperreal credences. *The Philosophical Review*. 123. 1. pp. 1–41.

Edgington, D. (1995) On conditionals. *Mind*. 104. 414. pp. 235–329.
Hájek, A. (2003) What conditional probability could not be. *Synthese*. 137. pp. 273–323.
Hájek, A. (2012) The fall of "Adams' Thesis"? *Journal of Logic, Language, and Information*. 21. 2. pp. 145–61.
Hájek, A. and Hall, N. (1994) The hypothesis of the conditional construal of conditional probability. In Eells, E. and Skyrms, B. (eds.), *Probability and Conditionals: Belief Revision and Rational Decision*. pp. 75–112. Cambridge: Cambridge University Press.
Halpern, J. (2010) Lexicographic probability, conditional probability, and nonstandard probability. *Games and Economic Behavior*. 68. 1. pp. 155–79.
Hill, B. M. (1980) On some statistical paradoxes and non-conglomerability. *Trabajos de estadística y de investigación operativa*. 31. 1. pp. 39–66.
Hill, B. M. and Lane, D. (1985) Conglomerability and countable additivity. *Sankhyā: The Indian Journal of Statistics, Series A*. 47. 3. pp. 366–79.
Humphreys, P. (1985) Why propensities cannot be probabilities. *The Philosophical Review*. 94. 4. pp. 557–70.
Humphreys, P. (2004) Some considerations on conditional chances. *British Journal for the Philosophy of Science*. 55. pp. 667–80.
Joyce, J. (1999) *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1986) Statistical implications of finitely additive probability. In Goel, P. K. and Zellner, A., (eds.) (1986) *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*. New York, NY: North-Holland.
Kolmogorov, A. N. (1933/1950) *Foundations of the Theory of Probability*. Translated from the German. New York, NY: Chelsea.
Levi, I. (1977) Subjunctives, dispositions, and chances. *Synthese*. 34. 4 pp. 423–55.
Lewis, D. (1976) Probabilities of conditionals and conditional probabilities. *The Philosophical Review*. 85. 3. pp. 297–315.
McCurdy, C. (1996) Humphreys's paradox and the interpretation of inverse conditional probabilities. *Synthese*. 108. pp. 105–25.
McGee, V. (1994) Learning the impossible. In Eells, E. and Skyrms, B. (eds.), *Probability and Conditionals*. Cambridge: Cambridge University Press.
Popper, K. (1955) Two autonomous axiom systems for the calculus of probabilities. *The British Journal for the Philosophy of Science*. 6. 21. pp. 51–7.
Popper, K. (1959) *The Logic of Scientific Discovery*, chapter iv*, pp. 326–48. New York, NY: Harper & Row.
Ramsey, F. P. (1931) Truth and probability. In Braithwaite, R. B., (ed.), *The Foundations of Mathematics and other Logical Essays*. pp. 156–98. New York, NY: Harcourt, Brace and Company.
Rényi, A. (1955) On a new axiomatic theory of probability. *Acta Mathematica Hungarica*. 6. 3. pp. 285–335.
Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (1984) The extent of non-conglomerability of finitely additive probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*. 66. pp. 205–26.
Seidenfeld, T., Schervish, M. J., and Kadane, J. B. (2014) Non-conglomerability for countably additive measures that are not κ-additive. Technical report. Carnegie Mellon University.
Stalnaker, R. (1970) Probability and conditionals. *Philosophy of Science*. 37. 1. pp. 64–80.
van Fraassen, B. (1995) Fine-grained opinion, probability, and the logic of full belief. *Journal of Philosophical Logic*. 24. pp. 349–77.

CHAPTER 10

THE BAYESIAN NETWORK STORY

RICHARD E. NEAPOLITAN AND XIA JIANG

Bayesian networks are now among the leading architectures for reasoning with uncertainty in artificial intelligence. This chapter concerns their story, namely what they are, how and why they came into being, how we obtain them, and what they actually represent. Each of these topics is discussed in turn.

10.1 What is a Bayesian Network?

Long before the arrival of Bayesian networks, Bayes' Theorem was used to perform probabilistic inference with two related random variables. The following example provides an illustration.

Example 1 *Suppose Dave plans to marry, and to obtain a marriage licence in the state in which he resides, one must take the blood-test enzyme-linked immunosorbent assay (ELISA), which tests for the presence of human immunodeficiency virus (HIV). Dave takes the test and it comes back positive for HIV. How likely is it that Dave is infected with HIV? Without knowing the accuracy of the test, Dave really has no way of knowing how probable it is that he is infected with HIV.*

The data we ordinarily have on such tests are the true positive rate (sensitivity) and the true negative rate (specificity). The true positive rate is the number of people who both have the infection and test positive divided by the total number of people who have the infection. For example, to obtain this number for ELISA, 10,000 *people who were known to be infected with HIV were identified. This was done using the Western Blot, which is the gold standard test for HIV. These people were then tested with ELISA, and* 9990 *tested positive. Therefore, the true positive rate is* 0.999. *The true negative rate is the number of people who both do not have the infection and test negative divided by the total number of people who do not have the infection. To obtain this number for ELISA* 10,000 *nuns who denied risk factors for HIV infection were tested. Of these,* 9980 *tested negative using the ELISA test. Furthermore, the* 20 *positive-testing nuns tested negative using the Western Blot test. So, the true negative rate is*

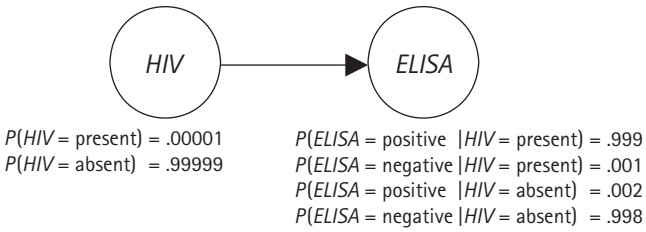


FIGURE 10.1 A two-node Bayesian network.

0.998, *which means that the false positive rate is* 0.002. *We therefore formulate the following random variables and probabilities:*

$$P(ELISA = positive|HIV = present) = 0.999$$
$$P(ELISA = positive|HIV = absent) = 0.002.$$

Note that neither of these probabilities is the probability we need, namely the probability of HIV being present given that someone tests positive. To obtain this probability, we also need the prior probability of HIV being present, and then an application of Bayes' Theorem. Recall that Dave took the blood test simply because the state required it. He did not take it because he thought for any reason he was infected with HIV. So, the only other information we have about Dave is that he is a male in the state in which he resides. Therefore if 1 *in* 100,000 *men in Dave's state is infected with HIV, we assign the following probability:*

$$P(HIV = present) = 0.00001.$$

We now employ Bayes' Theorem to compute

$$\begin{aligned} &P(present|positive) \\ &= \frac{P(positive|present)P(present)}{P(positive|present)P(present) + P(positive|absent)P(absent)} \\ &= \frac{(.999)(.00001)}{(.999)(.00001) + (.002)(.99999)} \\ &= 0.00497. \end{aligned}$$

Surprisingly, we are fairly confident that Dave is not infected with HIV. This is owing to the small prior probability.

We can summarize the information used in the previous application of Bayes' Theorem in Figure 10.1, which is a two-node/variable Bayesian network. Notice that it represents the random variables *HIV* and *ELISA* by nodes in a directed acyclic graph (DAG) and the causal relationship between these variables with an edge from *HIV* to *ELISA*. That is, the presence of *HIV* has a causal effect on whether the test result is positive; so there is an edge from *HIV* to *ELISA*. Besides showing a DAG representing the causal relationships, Figure 10.1 shows the prior probability distribution of *HIV* and the conditional probability distribution of *ELISA* given each value of its parent *HIV*. In general, a Bayesian network consists of

a DAG, whose edges represent relationships among random variables that are often (but not always) causal; the prior probability distribution of every variable that is a root in the DAG; and the conditional probability distribution of every non-root variable given each set of values of its parents. We use the terms *node* and *variable* interchangeably in discussing Bayesian networks.

Let's illustrate a more complex Bayesian network by considering the problem of detecting credit-card fraud (taken from Heckerman, 1996). Suppose that we have identified the following variables as being relevant to the problem:

Variable	What the Variable Represents
Fraud (*F*)	Whether the current purchase is fraudulent
Gas (*G*)	Whether gas has been purchased in the last 24 hours
Jewelry (*J*)	Whether jewelry has been purchased in the last 24 hours
Age (*A*)	Age of the card holder
Sex (*S*)	Sex of the card holder

These variables are all causally related. That is, a credit-card thief is likely to buy gas and jewelry, and middle-aged women are most likely to buy jewelry, whereas young men are least likely to buy jewelry. Figure 10.2 shows a DAG representing these causal relationships. Notice that it also shows the conditional probability distribution of every non-root variable given each set of values of its parents. The Jewelry variable has three parents, and there is a conditional probability distribution for every combination of values of those parents. The DAG and the conditional distributions together constitute a Bayesian network.

We used Bayes' Theorem to compute $P(HIV = present|ELISA = positive)$ from the information in the Bayesian network in Figure 10.1. Similarly, we can compute the probability of credit-card fraud given values of the other variables in the Bayesian network in Figure 10.2. For example, we can compute $P(F = yes|G = yes, J = yes, A =< 30, S = female)$. If this probability is sufficiently high, we can deny the current purchase or require additional identification. The computation is not a simple application of Bayes' Theorem as was the case for the two-node Bayesian network in Figure 10.1. Rather it is done using sophisticated inference algorithms (Neapolitan, 1989, 2004; Pearl, 1988).

Now that we have illustrated Bayesian networks, we define them formally. First, some notation is provided. Suppose we have a joint probability distribution P of the random variables X, Y, and Z, and X and Y are *conditionally independent* given Z. That is, for all values of X, Y, and Z

$$P(X|Y,Z) = P(X|Z).$$

Then we write

$$I_P(X, Y|Z).$$

In general, X, Y, and Z can represent sets of random variables.

Definition 1 *Suppose we have a joint probability distribution P of the random variables in some set* V *and a directed acyclic graph (DAG)* $\mathbb{G} = (\mathrm{V},\mathrm{E})$. *We say that* $(\mathbb{G},P)$ *satisfies the*

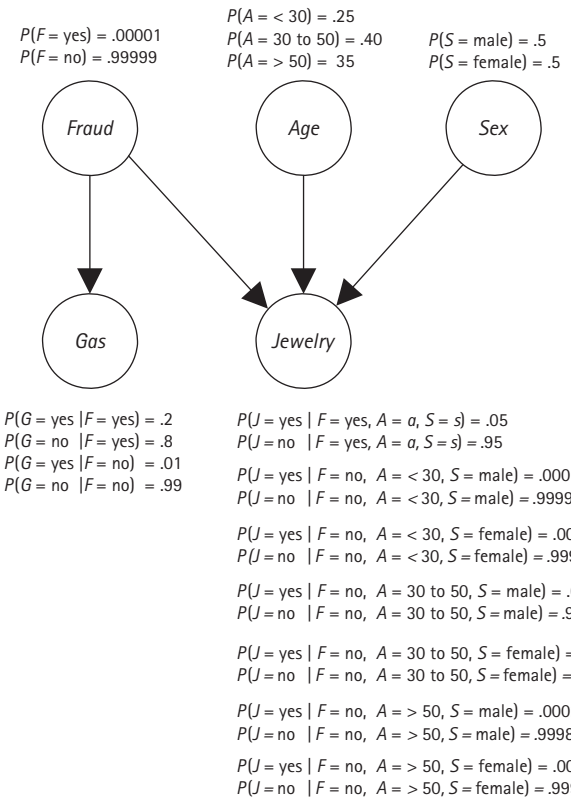


FIGURE 10.2 Bayesian network for detecting credit-card fraud. The notation $A = a$ means for all values a of the random variable A.

Markov condition if for each variable $X \in \mathrm{V}$, X is conditionally independent of the set of all its nondescendents given the set of all its parents. This means that if we denote the sets of parents and nondescendents of X by PA *and* ND*, respectively, then*

$$I_P(X, \mathrm{ND}|\mathrm{PA}).$$

If $(\mathbb{G}, P)$ satisfies the Markov condition, $(\mathbb{G}, P)$ is called a Bayesian network.

In a DAG by a *descendent* of X we mean a node Y such that there is a directed path from X to Y. A *nondescendent* of X is a node that is not equal to X and is not a descendent of X.

A Bayesian network $(\mathbb{G}, P)$, by definition, is a DAG $\mathbb{G}$ and joint probability distribution P that together satisfy the Markov condition. Then why in Figures 10.1 and 10.2 do we show a Bayesian network as a DAG and a set of conditional probability distributions? The reason is that $(\mathbb{G}, P)$ satisfies the Markov condition if and only if P is equal to the product of its conditional distributions in $\mathbb{G}$. Specifically, we have the following theorem.

Theorem 1 *$(\mathbb{G}, P)$ satisfies the Markov condition (and therefore is a Bayesian network) if and only if P is equal to the product of its conditional distributions of all nodes given their parents in G, whenever these conditional distributions exist.*

Proof. *The proof can be found in Neapolitan (2004).* ■

Example 2 *Consider the DAG in Figure 10.2. According to this theorem if our probability distribution P is defined by this product*

$$P(G,J,F,A,S) = P(G|F)P(J|F,A,S)P(F)P(A)P(S),$$

then (G,P) is a Bayesian network. Furthermore, if (G,P) is a Bayesian network, then P must be equal to the preceding product.

Owing to Theorem 1, we can represent a Bayesian network $(\mathbb{G}, P)$ using the DAG $\mathbb{G}$ and the conditional distributions. We don't need to show every value in the joint distributions. These values can all be computed from the conditional distributions. So we always show a Bayesian network as the DAG and the conditional distributions as we did in Figures 10.1 and 10.2. Herein lies the representational power of Bayesian networks. If there is a large number of variables, there are many values in the joint distribution. However, if the DAG is sparse, there are relatively few values in the conditional distributions. For example, suppose all variables are binary, there are 100 variables, and a joint distribution satisfies the Markov condition with a DAG such that each variable in the DAG has at most two parents. Then there are 2^{100} values in the joint distribution and fewer than 800 values in the conditional distributions. We see that a Bayesian network is a structure for representing a joint probability distribution succinctly.

We stress that we can't take just any DAG and expect a joint distribution to equal the product of its conditional distributions in the DAG. This is true only if the Markov condition is satisfied. It seems that we are left in a dilemma. That is, our goal is to succinctly represent a joint probability distribution using a DAG and conditional distributions for the DAG (a Bayesian network) rather than enumerating every value in the joint distribution. However, we don't know which DAG to use until we check whether the Markov condition is satisfied, and, in general, we would need to have the joint distribution to check this. A common way to resolve this dilemma is to construct a *causal DAG*, which is a DAG in which there is an edge from X to Y if and only if X is a direct cause of Y. The DAGs in Figures 10.1 and 10.2 are causal. We discuss further why a causal DAG should satisfy the Markov condition with the probability distribution of the variables in the DAG in Section 10.3.

10.2 The Genesis of Bayesian Networks

Early efforts in artificial intelligence were directed toward the development of all-purpose intelligent programs, which worked in limited domains and solved relatively simple problems. However, these programs failed to scale up so that they could handle difficult problems. In the 1970s many researchers turned their efforts to developing useful systems that solved difficult problems in specialized domains. These systems used powerful,

domain-specific knowledge and are called *knowledge-based systems* or *expert systems*. Initially, the knowledge was represented by rules about a particular domain, and the reasoning consisted of general-purpose algorithms that manipulated the rules. Successful knowledge-based systems include DENDRAL (Lindsay et al., 1980), a system that analyzes mass spectrograms in chemistry, XCON (McDermott, 1982), a system that configures VAX computers, and ACRONYM (Brooks, 1981), a vision support system. All these systems performed certain (logical) inference.

Realizing that in many domains (e.g. medical diagnosis) we cannot be certain of our conclusions, researchers searched for ways to incorporate uncertainty in the rules in their knowledge-based systems. The most notable such effort was the incorporation of *certainty factors* in the MYCIN system (Buchanan and Shortliffe, 1984). MYCIN is a medical expert system for diagnosing bacterial infections and prescribing treatments for them. Neapolitan (1989) showed that the rule-based representation of uncertain knowledge and reasoning not only is cumbersome and complex, but also does not model very well how humans reason. Pearl (1986) made the more reasonable conjecture that humans identify local probabilistic causal relationships between individual propositions and reason with these relationships. At this same time researchers in decision analysis (Shachter, 1986) were developing influence diagrams, which provide us with normative decisions in the face of uncertainty. In the 1980s researchers from cognitive science (e.g. Judea Pearl), computer science (e.g. Peter Cheeseman and Lotfi Zadeh), decision analysis (e.g. Ross Shachter), medicine (e.g. David Heckerman and Gregory Cooper), mathematics and statistics (e.g. David Spiegelhalter and Richard Neapolitan), and philosophy (e.g. Henry Kyburg) met at the newly formed Workshop on Uncertainty in Artificial Intelligence (now a conference) to discuss how to best perform uncertain inference in artificial intelligence. The texts *Probabilistic Reasoning in Intelligent Systems* (Pearl, 1988) and *Probabilistic Reasoning in Expert Systems* (Neapolitan, 1989) integrated many of the results of these discussions into the field we now call Bayesian networks. Bayesian networks have arguably become the standard for handling uncertain inference in AI, and many AI applications have been developed using them. Neapolitan and Jiang (2012) discuss many of them.

10.3 Learning Bayesian Networks

Once we know the structure (DAG) in a Bayesian network, learning the parameters (conditional probability distributions) is fairly straightforward. Each parameter can be estimated from data in the same way any probability distribution is estimated from data. Alternatively, we can ask an expert to estimate the parameters.

Learning the structure is more difficult; we address this matter next.

10.3.1 Building the Structure by Hand

Until the early 1990s the DAG in a Bayesian network was ordinarily hand-constructed by a domain expert. Then the conditional probabilities were assessed by the expert, learned from data, or obtained using a combination of both techniques. These DAGs were ordinarily constructed by placing an edge from node X to node Y if and only if X was a direct cause of Y. The resultant DAG is called a *causal DAG*. Next we discuss why a causal DAG should

satisfy the Markov condition with the joint probability distribution of the variables in the DAG.

First we provide an *operational definition* of a cause. We say we *manipulate* X when we force X to take some value, and we say X *causes* Y if there is some manipulation of X that leads to a change in the probability distribution of Y. A manipulation consists of a *randomized controlled experiment (RCE)* using some specific population of entities (e.g. individuals with chest pain) in some specific context (e.g. they currently receive no chest pain medication and they live in a particular geographical area). The causal relationship discovered is then relative to this population and this context. A *causal graph* is a directed graph containing a set of causally related random variables V such that for every $X, Y \in \mathrm{V}$ there is an edge from X to Y if and only if X is a cause of Y, and there is no subset of variables W of V such that if we knew the values of the variables in W, a manipulation of X would no longer change the probability distribution of Y. If there is an edge from X to Y, we call X a *direct cause* of Y. Note that whether or not X is a direct cause of Y depends on the variables included in V. A causal graph is a *causal DAG* if the causal graph is acyclic (i.e. there are no causal feedback loops).

If we assume that the observed probability distribution P of a set of random variables V satisfies the Markov condition with the causal DAG $\mathbb{G}$ containing the variables, we say we are making the *causal Markov assumption*, and we call $(\mathbb{G}, P)$ a *causal network*.

The next example illustrates why researchers feel that the Markov assumption is reasonable.

Example 3 *A history of smoking (H) is known to cause both bronchitis (B) and lung cancer (L). Lung cancer and bronchitis both cause fatigue (F), but only lung cancer can cause a chest X-ray (X) to be positive. There are no other causal relationships among the variables. Figure 10.3 shows a causal DAG containing these variables. The causal Markov assumption for that DAG entails the following conditional independencies.*

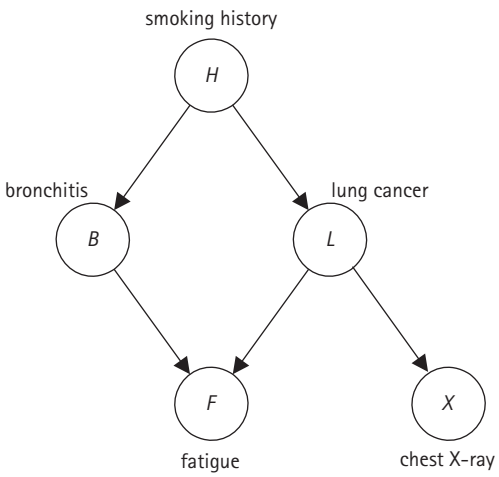


FIGURE 10.3 A causal DAG.

Node	Parents	Nondescendents	Conditional Independency
H	$\varnothing$	$\varnothing$	None
B	H	L, X	$I_P(B, \{L, X\} \mid H)$
L	H	B	$I_P(L, B \mid H)$
F	B, L	H, X	$I_P(F, \{H, X\} \mid \{B, L\})$
X	L	H, B, F	$I_P(X, \{H, B, F\} \mid L)$

Given the causal relationship in Figure 10.3, we would not expect bronchitis and lung cancer to be independent, because if someone had lung cancer it would make it more probable that the individual smoked (since smoking can cause lung cancer), which would make it more probable that another effect of smoking, namely bronchitis, was present. However, if we knew someone smoked, it would already be more probable that the person had bronchitis. Learning that the individual had lung cancer could no longer increase the probability of smoking (which is now 1), which means it cannot change the probability of bronchitis. That is, the variable H shields B from the influence of L, which is what the causal Markov condition says. Similarly, a positive chest X-ray increases the probability of lung cancer, which in turn increases the probability of smoking, which in turn increases the probability of bronchitis. So, a chest X-ray and bronchitis are not independent. However, if we knew the person had lung cancer, the chest X-ray could not change the probability of lung cancer and thereby change the probability of bronchitis. So B is independent of X conditional on L, which is what the causal Markov condition says.

There are three situations in which a causal graph should not satisfy the Markov condition. The first one is when there is a causal feedback loop. For example, perhaps studying causes good grades and good grades cause a student to study harder. When there is a causal feedback loop, our graph is not even a DAG.

The second situation is when a hidden common cause is present. The following example illustrates the problem with hidden common causes.

Example 4 *Suppose we wanted to create a causal DAG containing the variables cold (C), sneezing (S), and runny nose (R). Since a cold can cause both sneezing and a runny nose and neither of these conditions can cause each other, we would create the DAG in Figure 10.4 (a). The causal Markov condition for that DAG would entail $I_P(S, R \mid C)$. However, if there were a hidden common cause of S and R as depicted in Figure 10.4 (b), this conditional independency would not hold because even if the value of C were known, S would change the probability of H, which in turn would change the probability of R. Indeed, there is at least one other cause of sneezing and runny nose, namely hay fever. So when making the causal Markov assumption, we must be certain that we have identified all common causes.*

The final situation is more subtle. It concerns the presence of *selection bias*. The following example illustrates this situation.

Example 5 *The pharmaceutical company Merck had been marketing its drug finasteride as medication for men with benign prostatic hyperplasia (BPH). Based on anecdotal evidence, it seemed that there was a correlation between use of the drug and regrowth of scalp hair. Let's assume that Merck took a random sample from the population of interest and, based*

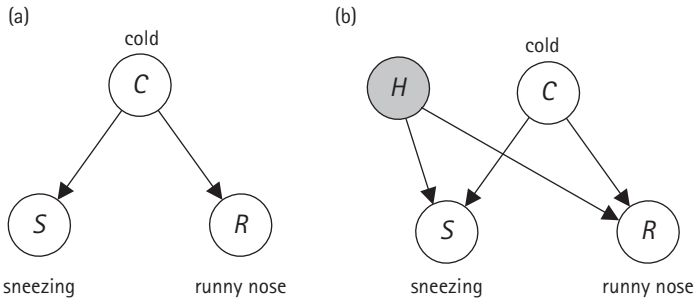


FIGURE 10.4 The causal Markov assumption would not hold for the DAG in (a) if there is a hidden common cause as depicted in (b).

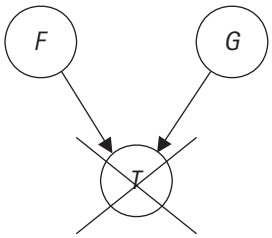


FIGURE 10.5 T is instantiated.

on that sample, determined that there is a correlation between finasteride use and hair regrowth. Assume further that there could be no hidden common causes of finasteride use and hair regrowth. Should Merck conclude that finasteride causes hair regrowth and therefore market it as a cure for baldness? Not necessarily. There is yet another possible causal explanation for this correlation. Suppose our sample (or even our entire population) consists of individuals who have some (possibly hidden) effect of both finasteride and hair regrowth. For example, suppose finasteride (F) and apprehension about lack of hair regrowth (G) both cause hypertension,[1] *and our sample consists of individuals who have hypertension (T). We say a node is instantiated when we know its value for the entity currently being modeled. So we are saying the variable T is instantiated to the same value for every entity in our sample. This situation is depicted in Figure 10.5, where the cross through T means that the variable is instantiated. Usually, the instantiation of a common effect creates a dependency between its causes because each cause explains the occurrence of the effect, thereby making the other cause less likely. Psychologists call this discounting. So, if this were the case, discounting would explain the correlation between F and G.*[2] *This type of dependency is called selection bias.*

[1] There is no evidence that either finasteride or apprehension about the lack of hair regrowth causes hypertension. This example is only for the sake of illustration.

[2] Merck eventually did a RCE involving 1879 men aged 18 to 41 with mild to moderate hair loss of the vertex and anterior mid-scalp areas. Half of the men were given 1 mg. of finasteride, whereas the other half were given 1 mg. of a placebo. The results indicated that finasteride does indeed cause hair regrowth. Merck now markets finasteride for hair regrowth under the label *propecia*.

10.3.2 Learning the Structure From Data

The second method for obtaining the DAG in a Bayesian network is to learn it from data. For example, suppose we have data concerning the values of the variables in Figure 10.3 on 10,000 individuals in some population. From these data we want to learn a structure like the DAG in Figure 10.3.

One way to do this is to *score* the DAG based on the data. A commonly used score is the probability of the data given each candidate DAG (Cooper and Herskovits, 1992). Other often used scores use the *Minimum Description Length* (*MDL*) *Principle* (Rissanen, 1978), which is based on information theory and says that the best model of a collection of data is the one that minimizes the sum of the encoding lengths of the data and the model itself. We do not discuss these scores here. You are referred to a text such as Neapolitan (2004) for an introduction to scoring methods.

A second, more intuitive method is called *constraint-based learning*; a brief introduction to this method to this method is provided next. In this approach, we try to learn a DAG from the conditional independencies in the generative probability distribution. We illustrate the constraint-based approach by showing how to learn a DAG when we assume the faithfulness condition, which is as follows.

Definition 2 *Suppose we have a joint probability distribution P of the random variables in some set* V *and a DAG* $\mathbb{G} = (\mathrm{V}, \mathrm{E})$. *We say that* $(\mathbb{G}, P)$ *satisfy the faithfulness condition if all and only the conditional independencies in P are entailed by the Markov condition in* $\mathbb{G}$.

The faithfulness condition assumes the Markov condition is satisfied, but goes beyond that condition by assuming that every conditional independency in P can be identified by applying the Markov condition to $\mathbb{G}$.

Example 6 *Suppose there are three random variables C, L, S, and the only conditional independency in their joint probability distribution P is that L and S are conditionally independency given C. That is,* $I_P(L, S|C)$. *Then P satisfies the Markov condition with the DAGs in Figure 10.6 (a) and Figure 10.6 (b), and P satisfies the faithfulness condition only with the DAG in Figure 10.6 (a).*

If we can find a DAG that is faithful to a probability distribution P, we have achieved our goal of representing P succinctly. That is, if there are DAGs faithful to P then those DAGs are the smallest DAGs that represent P (see Neapolitan, 2004). We say *DAGs* because if a DAG is faithful to P, then any Markov equivalent DAG is also faithful to P. Two DAGs are *Markov equivalent* if they entail the same conditional independencies. For example, the DAGs $L \rightarrow C \rightarrow S$ and $S \leftarrow C \leftarrow L$, which are Markov equivalent to the DAG $L \leftarrow C \rightarrow S$, are also faithful to the probability distribution P discussed in Example 6. As we shall see, not every probability distribution has a DAG that is faithful to it. However, if there are DAGs faithful to a probability distribution, it is relatively easy to discover them. We discuss learning a faithful DAG next.

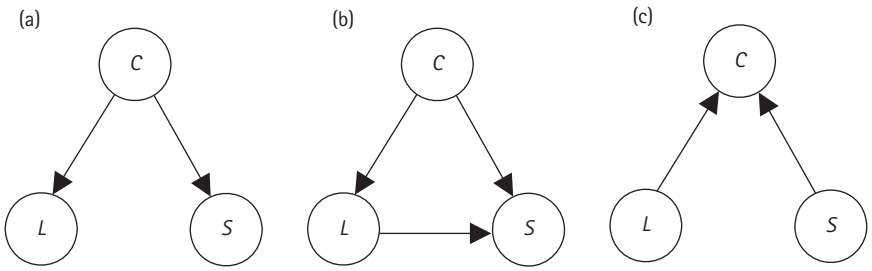


FIGURE 10.6 If the only conditional independency in P is $I_P(L,S|C)$, then P satisfies the Markov condition with the DAGs in (a) and (b), and P satisfies the faithfulness condition only with the DAG in (a).

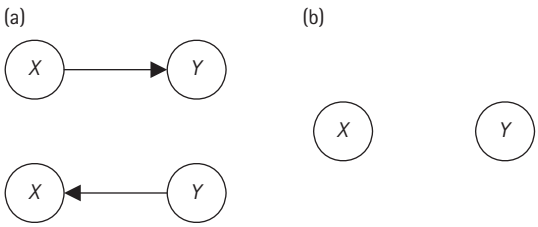


FIGURE 10.7 If the set of conditional independencies is $\{I_P(X,Y)\}$, we must have the DAG in (b), whereas if it is $\varnothing$, we must have one of the DAGs in (a).

10.3.2.1 *Learning a Faithful DAG*

We assume that we have a sample of entities from the population over which the random variables are defined, and we know the values of the variables of interest for the entities in the sample. The sample could be a random sample, or it could be obtained from passive data. From this sample, we have deduced the conditional independencies among the variables. A method for deducing conditional independencies and obtaining a measure of our confidence in them is described in Spirtes et al. (1993) and Neapolitan (2004). Our confidence in the DAG we learn is no greater than our confidence in these conditional independencies.

Next we give a sequence of examples in which we learn a DAG that is faithful to the probability distribution of interest. These examples illustrate how a faithful DAG can be learned from the conditional independencies if one exists. We stress again that the DAG is faithful to the conditional independencies we have learned from the data. We are not certain that these are the conditional independencies in the probability distribution for the entire population.

Example 7 *Suppose* V *is our set of observed variables,* $\mathrm{V} = \{X, Y\}$*, and the set of conditional independencies in P is*

$$\{I_P(X,Y)\}.$$

We want to find a DAG faithful to P. We cannot have either of the DAGs in Figure 10.7 (a). The reason is that these DAGs do not entail that X and Y are independent, which means the faithfulness condition is not satisfied. So, we must have the DAG in Figure 10.7 (b). We conclude that P is faithful to the DAG in Figure 10.7 (b).

Example 8 *Suppose* $\mathrm{V} = \{X, Y\}$ *and the set of conditional independencies in P is the empty set*

$$\varnothing.$$

That is, there are no independencies. We want to find a DAG faithful to P. We cannot have the DAG in Figure 10.7 (b). The reason is that this DAG entails that X and Y are independent, which means that the Markov condition is not satisfied. So, we must have one of the DAGs in Figure 10.7 (a). We conclude that P is faithful to both the DAGs in Figure 10.7 (a). Note that these DAGs are Markov equivalent.

Example 9 *Suppose* $\mathrm{V} = \{X, Y, Z\}$*, and the set of conditional independencies in P is*

$$\{I_P(X, Y)\}.$$

We want to find a DAG faithful to P. There can be no edge between X and Y in the DAG for the reason given in Example 7. Furthermore, there must be edges between X and Z and between Y and Z for the reason given in Example 8. We cannot have any of the DAGs in Figure 10.8 (a). The reason is that these DAGs entail $I_P(X, Y|Z)$*, and this conditional independency is not present. So, the Markov condition is not satisfied. Furthermore, the DAGs do not entail* $I_P(X, Y)$*. So, the DAG must be the one in Figure 10.8 (b). We conclude that P is faithful to the DAG in Figure 10.8 (b).*

Example 10 *Suppose* $\mathrm{V} = \{X, Y, Z\}$ *and the set of conditional independencies in P is*

$$\{I_P(X, Y|Z)\}.$$

We want to find a DAG faithful to P. For reasons similar to those given before, the only edges in the DAG must be between X and Z and between Y and Z. We cannot have the DAG in Figure 10.8 (b). The reason is that this DAG entails $I(X, Y)$*, and this conditional independency is not present. So, the Markov condition is not satisfied. So, we must have one of the DAGs in Figure 10.8 (a). We conclude that P is faithful to all the DAGs in Figure 10.8 (a).*

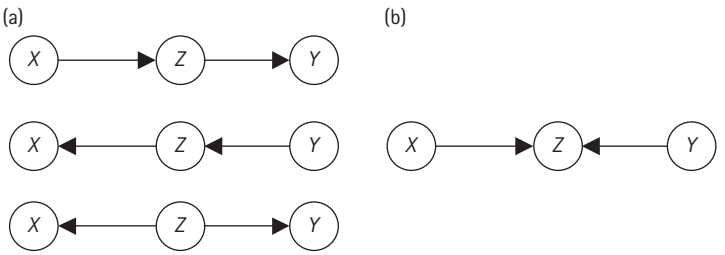


FIGURE 10.8 If the set of conditional independencies is $\{I_P(X, Y)\}$, we must have the DAG in (b); if it is $\{I_P(X, Y|Z)\}$, we must have one of the DAGs in (a).

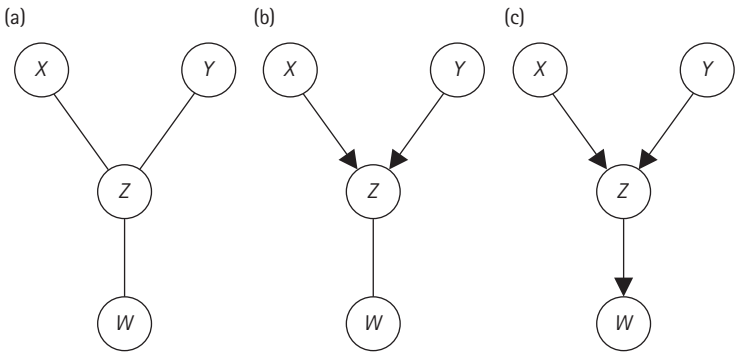


FIGURE 10.9 If the set of conditional independencies is $\{I_p(X,Y),\ I_P(W,\{X,Y\}|Z)\}$, we must have the DAG in (c).

We now state a theorem whose proof can be found in Neapolitan (2004). At this point your intuition should suggest that it is true.

Theorem 2 *If $(\mathbb{G},P)$ satisfies the faithfulness condition, then there is an edge between X and Y if and only if X and Y are not conditionally independent given any subset of the variables in $\mathbb{G}$.*

Example 11 *Suppose* $\mathrm{V}=\{X,Y,Z,W\}$ *and the set of conditional independencies in P is*

$$\{I_p(X,Y),\quad I_P(W,\{X,Y\}|Z)\}.$$

We want to find a DAG faithful to P. Owing to Theorem 2, the links (edges without regard for direction) must be as shown in Figure 10.9 (a). We must have the directed edges shown in Figure 10.9 (b) because we have $I_p(X,Y)$. Therefore, we must also have the directed edge shown in Figure 10.9 (c) because we do not have $I_P(W,X)$. We conclude that P is faithful to the DAG in Figure 10.9 (c).

Example 12 *Suppose* $\mathrm{V}=\{X,Y,Z,W\}$ *and the set of conditional independencies in P is*

$$\{I_P(X,\{Y,W\}),\quad I_P(Y,\{X,Z\})\}.$$

We want to find a DAG faithful to P. Owing to Theorem 2, we must have the links shown in Figure 10.10 (a). Now, if we have the chain $X\rightarrow Z\rightarrow W$, $X\leftarrow Z\leftarrow W$, or $X\leftarrow Z\rightarrow W$, then we do not have $I_P(X,W)$. So, we must have the chain $X\rightarrow Z\leftarrow W$. Similarly, we must have the chain $Y\rightarrow W\leftarrow Z$. So, our graph must be the one in Figure 10.10 (b). However, this graph is not a DAG. The problem here is that there is no DAG faithful to P.

Example 13 *Suppose we have the same vertices and conditional independencies as in Example 12. As shown in that example, there is no DAG faithful to P. However, this does not mean we cannot find a more succinct way to represent P than using a complete DAG. P satisfies the Markov condition with each of the DAGs in Figure 10.11. That is, the DAG in Figure 10.11 (a) entails*

$$\{I_P(X,Y),\quad I_P(Y,Z)\}$$

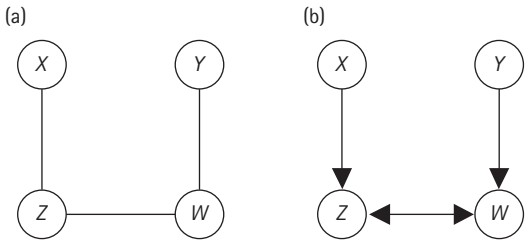


FIGURE 10.10 If the set of conditional independencies is $\{I_P(X,\{Y,W\}), I_P(Y,\{X,Z\})\}$ and we try to find a DAG faithful to P, we obtain the graph in (b), which is not a DAG.

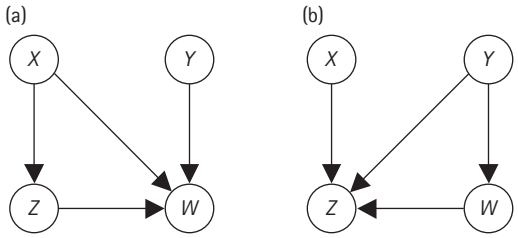


FIGURE 10.11 If the set of conditional independencies is $\{I_P(X,\{Y,W\}), I_P(Y,\{X,Z\})\}$, P satisfies the Markov condition with both these DAGs.

and these conditional independencies are both in P, whereas the DAG in Figure 10.11 (b) entails

$$\{I_P(X,Y), \quad I_P(X,W)\}$$

and these conditional independencies are both in P. However, P does not satisfy the faithfulness condition with either of these DAGs because the DAG in Figure 10.11 (a) does not entail $I_P(X,W)$*, whereas the DAG in Figure 10.11 (b) does not entail* $I_P(Y,Z)$.

Each of these DAGs is as succinct as we can make it in representing the probability distribution. So, when there is no DAG faithful to a probability distribution P, we can still represent P much more succinctly than we would by using the complete DAG. (A complete DAG is one in which there is an edge from every node to every other node.) A structure learning algorithm tries to find the most succinct representation. Depending on the number of possible values of each variable, one of the DAGs in Figure 10.11 may actually be a more succinct representation than the other because it contains fewer parameters. A constraint-based learning algorithm could not distinguish between the two, but a score-based one could. See Neapolitan (2004) for a complete discussion of this matter.

10.3.2.2 *Learning a DAG in Which P Is Embedded Faithfully*

In a sense we compromised in Example 13 because the DAG we learned did not entail all the conditional independencies in P. This is fine if our goal is to learn a Bayesian network that

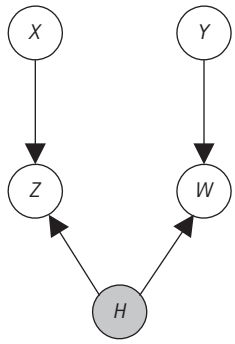


FIGURE 10.12 If the set of conditional independencies in P is $\{I_P(X,\{Y,W\}), I_P(Y,\{X,Z\})\}$, then P is embedded faithfully in this DAG.

will later be used to perform inference. However, another application of structure learning is *causal learning*, which is discussed in Spirtes et al. (1993) and Neapolitan (2004). When we're learning causes it would be better to find a DAG in which P is embedded faithfully. We discuss embedded faithfulness next.

Definition 3 *Suppose we have a joint probability distribution P of the random variables in some set* V *and a DAG* $\mathbb{G} = (\mathrm{W},\mathrm{E})$ *such that* $\mathrm{V} \subseteq \mathrm{W}$. *We say that* $(\mathbb{G},P)$ *satisfy the embedded faithfulness condition if all and only the conditional independencies in P are entailed by the Markov condition in* $\mathbb{G}$, *restricted to variables in* V. *Furthermore, we say that P is embedded faithfully in* $\mathbb{G}$.

Example 14 *Again suppose* $\mathrm{V} = \{X,Y,Z,W\}$ *and the set of conditional independencies in P is*

$$\{I_P(X,\{Y,W\}), \quad I_P(Y,\{X,Z\})\}.$$

Then P is embedded faithfully in the DAG in Figure 10.12. By including the hidden variable H in the DAG, we are able to entail all and only the conditional independencies in P restricted to variables in V.

10.4 What Probability Distribution is Represented?

The *relative frequency approach to probability* (von Mises, 1919) concerns repeatable experiments such as tossing a coin, or sampling individuals and determining whether they smoke. In this approach the probability of an outcome is defined to be the limit as the number of trials approaches infinity of the relative frequency with which the outcome occurs. Such probabilities are often called *relative frequencies*.[3] We can approximate the

[3] Many authors such as Neapolitan (1989) note that in most applications it is not really tenable to assume that an exact relative frequency distribution exists in nature. However, our discussion will proceed assuming one does. The philosophical difference is not really important to this discussion.

true relative frequency (probability) by performing a large number of trials. For example, if we toss a thumbtack 10,000 times and it lands heads 3800 times,

$$P(\text{heads}) \approx \frac{3800}{10{,}000} = 0.38.$$

After the thumbtack lands heads 3800 times out of 10,000 tosses, we believe it has about a 0.38 chance of landing heads on the next toss, and we bet accordingly. That is, we would consider it fair to win about \$0.62 if the thumbtack landed heads and to lose $\$1 - \$0.62 = \$0.38$ if the thumbtack landed tails. Since the bet is considered fair, the opposite position, namely, to lose \$0.62 if the thumbtack landed heads and to win \$0.38 if it landed tails, would also be considered fair. Hence, we would take either side of the bet. This notion of a probability as a value that determines a fair bet is called a *subjective or Bayesian approach to probability* (Lindley, 1985), and probabilities assigned within this frame are called *beliefs*. It is applicable not only to repeatable experiments, but also to non-repeatable experiments such as a sporting event or the stock markets performance on a particular day. For example, one might assign 0.6 to the probability that the football team the Chicago Bears will beat the Buffalo Bills in their September 7, 2014 season opener and bet accordingly.

The probability distribution in a Bayesian network is ordinarily based on the notion of a probability as a relative frequency, not one that is solely a belief such as the probability of the Bears beating the Bills. However, assuming that there is an actual relative frequency distribution of the random variables in nature, is the distribution represented by the Bayesian network this distribution? For example, does the Bayesian network in Figure 10.2 contain the actual relative frequency distribution of the variables in that network? If not, what probability distribution does it contain?

To investigate these questions, consider the following two ways of the developing a Bayesian network:

1. Suppose we have the five random variables in Figure 10.3, which are a history of smoking (H), bronchitis (B), lung cancer (L), fatigue (F), and a chest X-ray (X). Owing to Theorem 1, if we create an arbitrary DAG $\mathbb{G}$ containing these variables and learn the values of the conditional distributions in $\mathbb{G}$ from a very large amount of data (so that we are essentially certain of the values), then the product P of those conditional distributions will satisfy the Markov condition with $\mathbb{G}$, which means $(\mathbb{G}, P)$ is a Bayesian network. However, the actual relative frequency distribution P' of the random variables will not satisfy the Markov condition with an arbitrary DAG, which means that P is not even an approximation of the actual distribution P'.
2. Now suppose we have the same five random variables, we know the conditional independencies in the actual relative frequency distribution P', we construct a DAG $\mathbb{G}$ that contains these conditional independencies, and we again learn the values of the conditional distributions in $\mathbb{G}$ from a very large amount of data. In this case the product P of these conditional distributions will be essentially the same as P'.

In practice the Bayesian networks we develop are in between these two extremes. We develop a DAG that we believe comes close to representing the conditional independencies (either from expert judgment or data), and then we assign values to the conditional distributions, which are ordinarily obtained from reasonably sized datasets and which therefore are

estimates of the actual probabilities (relative frequencies). What distribution is represented by the resultant Bayesian network? It is our joint subjective probability distribution *P* of the variables obtained from our beliefs concerning conditional independencies among the variables (the structure of the DAG $\mathbb{G}$) and our data. Note that if we are correct about the conditional independencies, we will have convergence (as the size of our sample approaches infinity) to the actual relative frequency distribution.

Acknowledgments

This work was supported by National Library of Medicine grants number R00LM010822, R01LM011663, and R01LM011962.

References

Brooks, R. A. (1981) Symbolic Reasoning Among 3-D Models and 3-D Images. *Artificial Intelligence*. 17.

Buchanan, B. G., and Shortliffe, E. H. (1984) *Rule-Based Expert Systems*. Reading, MA: Addison-Wesley.

Cooper, G. F. and Herskovits, E. (1992) A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*. 9.

Heckerman, D. (1996) A Tutorial on Learning with Bayesian Networks. Technical Report # MSR-TR-95-06. Microsoft Research.

Lindley, D. V. (1985) *Introduction to Probability and Statistics from a Bayesian Viewpoint*. London: Cambridge University Press.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A. and Lederberg, J. (1980) *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project*. USA: McGraw-Hill.

McDermott, J. (1982) A Rule-Based Configurer of Computer Systems. *Artificial Intelligence*. 19. 1.

Neapolitan, R. E. (1989) *Probabilistic Reasoning in Expert Systems*. New York, NY: Wiley.

Neapolitan, R. E. (2004) *Learning Bayesian Networks*. Upper Saddle River, NJ: Prentice Hall.

Neapolitan, R. E. and Jiang, X. (2012) *Contemporary Artificial Intelligence*. Boca Raton, FL: CRC Press.

Pearl, J. (1986) Fusion, Propagation, and Structuring in Belief Networks. *Artificial Intelligence*. 29.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. San Francisco, CA: Morgan Kaufmann.

Rissanen, J. (1978) Modeling by Shortest Data Description. *Automatica*. 14.

Shachter, R. D. (1986) Evaluating Influence Diagrams. *Operations Research*. 34.

Spirtes, P., Glymour, C., and Scheines, R. (1993) *Causation, Prediction, and Search*. New York, NY: Springer-Verlag. (2nd ed. (2000) Cambridge, MA: MIT Press).

von Mises, R. (1919) Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*. 5.

PART III

ALTERNATIVES TO STANDARD PROBABILITY THEORY

CHAPTER 11

MATHEMATICAL ALTERNATIVES TO STANDARD PROBABILITY THAT PROVIDE SELECTABLE DEGREES OF PRECISION

TERRENCE L. FINE

11.1 Prologue

There is a firmly and widely held assumption that standard mathematical probability has only one meaning (although there is active disagreement as to which it is of two possibilities) and that the only mathematical representation of chance, uncertain, or indeterminate phenomena must be through real-valued probability, as described in Section 11.2. One might say, recalling Bertrand Russell, that an exclusive reliance upon standard probability, "....has many advantages; they are the same as the advantages of theft over honest toil." "Theft" corresponding to seeking comfort in recourse to familiar manipulations of numerically-valued quantities and functions at the expense of the "toil" of developing mathematical approaches that more faithfully represent the phenomena they are meant to describe.

While much philosophical literature refers only to earlier philosophical literature, in this instance I need you to reconsider certain familiar real-world situations that involve chance, uncertain, or indeterminate phenomena. We are supported in this by von Neumann.

> As a mathematical discipline travels far from its empirical source....it is beset with very grave dangers....that the subject will develop along the line of least resistance, that the stream, so far from its source, will separate into a multitude of insignificant branches….
>
> In any event, whenever this state is reached, the only remedy seems to me to be the rejuvenating return to the source: the re-injection of more or less directly empirical ideas. (von Neumann 1947, part 2, p. 4)

In Section 11.3 we discuss four such real-world situations. Our two goals are first to make credible the shortcomings of relying solely upon standard mathematical probability

and second to motivate your interest in particular alternative approaches to mathematical characterizations of chance, uncertain, and indeterminate phenomena.

Section 11.4.1 discusses issues of modelling accuracy and precision in an *empirical domain*[1] in which issues of chance, uncertainty, or indeterminacy are central. This argument is further developed in the context of frequentist (see Section 11.4.2) and subjectivist (see Section 11.4.3) approaches. In Sections 11.4.4 and 11.4.5 we proceed more abstractly to examine the nature of representation in a mathematical domain of those aspects of an empirical domain that involve kinds of chance, uncertainty, or indeterminacy. We employ the viewpoint of *measurement theory* (not to be confused with measure theory), as largely developed in the social sciences, so as to enable us to better understand the relationship between the empirical and the mathematical domains. The discussions in Sections 11.3 and 11.4 expose the limitations of a sole reliance upon standard probability.

Sections 11.5 through 11.8 outline the following viable mathematical alternatives to standard probability, as arranged in order of increasing precision:

1. Modal/classificatory probability;
2. Comparative probability;
3. (a) Belief and plausibility functions;
   (b) Upper and lower probability.

Section 11.9 makes use of the upper and lower probability models introduced in Section 11.8 to make probability-like models that do not prejudge issues of convergence of time averages like relative frequencies when our only data is of finite duration. The lower probability models considered in Section 11.9 will use familiar standard probability modeling concepts of stationarity and bounded random variables without then automatically committing us to assertions of convergent averages of such terms that are forced by certain theorems of standard probability. Issues of convergence that are settled by the mathematical constraints forced by the axioms of standard probability are better left to be resolved by practitioners of the sciences concerned with the phenomena being modeled. In this connection, see also the quotation from John Venn in Section 11.4.2 about the inevitably unstable very long-run behavior of relative frequencies.

In focusing on alternatives to unconditional standard probability, I omit discussion of a large number of important topics, especially conditioning, but also computational complexity-based probability, and the various meanings of probability concepts. The thrust of our argument is against probability always having to have a unique real value, and this argument is best advanced without considering the additional issues raised by, say, updating or conditioning.

[1] Several terms of special importance and/or having technical meaning are set in italics when first introduced. In our usage an empirical domain is a *small world* of empirical phenomena considered as detached from the much larger setting in which it is embedded. For example, in Section 11.3, only what impacts our car driving decisions in the next minute.

11.2 Unconditional Standard Probability

We are first exposed to *standard probability* at an early age, when our critical immune system is still forming. An elementary version is taught as so uncontroversial that there can be no more objection to it than to arithmetic. Unfortunately, there is rarely a later corrective offered for the limited scope of applicability of standard probability (but see Aidan Lyon in this volume; Hájek and Hitchcock 2016). This situation notwithstanding, in our daily lives we commonly use alternatives of very limited precision such as *probable*, "highly likely", "rarely", "unlikely", "I doubt it", or statements of odds (e.g.,"5 to 3").

Very compactly, *standard probability* was successfully axiomatized by Andrei N. Kolmogorov (1933/1956) (see Shafer and Vovk 2006 for a discussion) in terms of the following:

- *random experiment* $\mathcal{E} = (\Omega, \mathcal{A}, \mu)$ that is a triplet;
- the *sample space* $\Omega = \{\omega\}$ is the set of all *possible* (in some practical or physical sense) outcomes of the performance of $\mathcal{E}$ and is called the *sure event*;
- a collection $\mathcal{A}$ of subsets of Ω that are of interest, and about which we have knowledge concerning their tendency to occur or whether or not they are true;
- $\mathcal{A}$ is required to be an σ*-algebra*[2]
- a *real-valued function* μ[3], defined for each set in $\mathcal{A}$, that is *non-negative*, assigns $\mu(\Omega) = 1$ (*unit normalization*), and is *countably additive* for *disjoint sets*[4].

Finite additivity is defined for two events A, B by

$$A \cap B = \emptyset \text{ implies } \mu(A \cup B) = \mu(A) + \mu(B).$$

Easy implications of these axioms include: the probability $\mu(A)$ of any event A is less than or equal to one; if a finite collection of events $A_1, A_2, \ldots, A_n$ is *pairwise disjoint* (for each $i \neq j$, $A_i \perp A_j$) then

$$\mu(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i).$$

The extension to n being infinite is either the *countable additivity axiom* of standard probability or the axiom of *monotone continuity*[5]. This latter axiom states that any sequence of events $\{A_i\}$ in the σ-algebra $\mathcal{A}$ that are nested (for all $i, A_i \supseteq A_{i+1}$) and that are shrinking to the $\emptyset$ ($\cap_{i=1}^{\infty} A_i = \emptyset$), satisfies $\lim_{i \to \infty} \mu(A_i) = 0$. The two axioms are equivalent given the other Kolmogorov axioms. Addition of either axiom has a consequence of ruling out an extension of the notion of a uniform/random distribution on a finite set Ω of outcomes to

[2] $\mathcal{A}$ is nonempty in that it must contain Ω and its complement Ω^c that is the empty set $\emptyset$. If A, B are in $\mathcal{A}$ then so are their union $A \cup B$ and intersection $A \cap B$. When Ω is an infinite set, then standard probability requires $\mathcal{A}$ to be a σ-algebra, a collection closed under countable unions and intersections.

[3] We will use μ throughout to denote a standard probability measure and reserve P for alternative probability ideas.

[4] Disjoint sets have no elements in common, sometimes denoted using $A \perp B$ for A, B disjoint.

[5] A restricted version of monotone continuity will be needed in Section 11.9.8.

the case of countably infinite Ω. When Ω is an infinite set and we have infinitely many events of interest then $\mathcal{A}$ is a *σ-algebra of events* (a collection closed under set complementation, countable unions, and hence countable intersections).

The above axioms are augmented by the definitions of expectation, conditional probability, and the key concept of independence. This chapter, however, will focus only on the counterparts to probability.

It would be remiss not to credit de Finetti (1990) for providing an extensive and much more fully reasoned development of mathematical probability that explores with care questions of additivity. The extraordinarily influential treatment by Kolmogorov (1933/1956) established the rigorous and general mathematical foundations of standard probability, as well as providing many of the basic theorems of this subject.

11.3 Familiar Instances Where Probabilistic Reasoning Places Little Reliance Upon Standard Probability

Experiences with many successful applications (e.g., in the physical and social sciences and games of chance) and the lazy belief that "it works" in a wide range of contexts, in which its applicability has not always been critically examined, combine to support the widespread acceptance and use of standard probability. Regarding the weak defense of "it works", there are instances (e.g., in Gaussian models of economic time series) of adherence to specific standard probability models in the face of evidence (10σ exceedances) that strongly contradict these accepted models. Given the difficulty encountered in accepting the occasional inapplicability of such familiar specific standard probability models, it is to be expected that it will be much more difficult to accept the occasional inapplicability of well-entrenched standard probability itself.

11.3.1 Setting and Narrative Motivation

A reasoner (not necessarily human, often a single individual, sometimes a group of individuals) R engages in probabilistic reasoning at initial time t_i to make a decision or inference at a later time t_f. In order to do so, R needs *information* relevant to the specific small world (the environs of the question at hand) or *empirical domain* (see Section 11.4.5) of chance, uncertain, indeterminate, etc. events that bears on the inference or decision to be made at t_f. Some of this information was available to R prior to t_i. The remainder of the information that R may use is obtained during the inference process time interval from t_i to t_f. The total information under consideration by R prior to time t_f is denoted by $\mathcal{I}_{R,t_i,t_f}$.[6] Prior to t_f, R will have used $\mathcal{I}_{R,t_i,t_f}$ to choose a probability-like model, perhaps of the types

[6] A formal concept of knowledge, denoted by $\mathcal{K}_{R,t_f}$, was developed and used by Kyburg (1974), chapter 8, and Levi (1985) for *rational corpora*.

discussed in this chapter. The chosen probability-like model then directs the decision or inference adopted at t_f. In order to gain a vantage point, and motivate your interest in alternatives, we offer four real-world narratives that illustrate how wide a stretch it can be to attempt to resolve the issues raised in these narratives by recourse to standard probability.

11.3.2 Driving a Car

11.3.2.1 *How People Do This*

In this narrative the driver is the individual R (see yourself here), and the inference is about a safe road speed for a stretch of road ahead. Inferences need to be made rapidly about the choice of a safe speed, say, for an upcoming turn in a quarter-mile on a wintry road at dusk, that is first sighted at time t_i and entered at time t_f. $\mathcal{I}_{R,t_i,t_f}$ assembled by R, describes an empirical domain relevant to the trajectory of a specific car traveling on a specific section of road and guided by R. $\mathcal{I}_{R,t_i,t_f}$ has a mix of objective and subjective elements. Elements prior to t_i might include: the driver's familiarity with this car and similar ones and with this road and similar ones in a variety of weather conditions; earlier road conditions; time of day, visibility, and weather conditions; the rough consequences to R of various scenarios. Elements assembled between t_i and t_f might include: behavior of cars up ahead (e.g., skidding, speed); response of the car to cautious, probing steering changes; and, if applicable, the solicited opinions of passengers. The $\mathcal{I}_{R,t_i,t_f}$ assembled by the driver R might provide support for assertions such as: "slowing rapidly is quite likely to lead to an accident"; "slowing gently is much less likely to lead to an accident in the turn than is maintaining speed".

Our own introspection, and familiarity with the behavior of others, reveals that Driver R, in the allotted time of $t_f - t_i$, is unlikely to derive and use numerical probabilities of any significant precision. A mathematical representation by, or a mental state interpretation in terms of, modal (Section 11.5) or comparative probability (Section 11.6), that makes no numerical assertions, better fits the decision problem facing R. Probability-like assessments of the outcomes of potential driver actions are not made to even low degrees of numerical precision. Indeed, rather than being explicitly numerical, the judgments of R may be qualitative and restricted to such alternatives as very unlikely, unlikely, likely, highly likely, or just comparisons of the likelihoods of the successes of a few decisions, or the elimination of certain alternatives (e.g., "slowing rapidly might lead to an uncontrolled skid").

11.3.2.2 *Bayesian Inference is Not Even for the Birds*

There are many who hold that we should compare such uncertain and risky alternatives using Bayesian methods (see Section 11.4.3) that require that R quickly represent numerically his/her degrees of belief of various eventualities and then speedily arrive at an assessment of the available actions. R requires substantial analytical powers, all being exercised over time periods of less than a quarter-minute at road speeds of 55mph. To the contrary, the animal kingdom provides evidence that this need not be the case. Our ability to discern and react to danger in a timely manner is something we share with birds and rodents. Studies (e.g., Emmerton c. 2002, Reznikova and Ryabko 2011) have shown these animals to have limited numeracy. The success of birds in reacting to dangers suggests that,

at best, very crude mathematical reasoning is necessary for their survival. It is plausible that we have evolved to survive everyday risks through robust reasoning that is neither mathematically sophisticated nor of a complexity comparable to such an account based upon numerical probabilities. As observed in Section 11.3.2.1, the lack of mathematical sophistication undergirding our survival argues against the applicability of some, but not all, of the alternative probability-like concepts to be addressed below.

11.3.2.3 *Google Self-Driving Cars*

Does the existence of the Google self-driving car contradict much of what I have said? Perhaps, depending upon the algorithm used and where, if anywhere, standard probability is used. As of the time of this writing, Google has yet to cope with winter weather involving snow and icy roads that impair traction and visibility and require looking well ahead,

> ...we'll need to master snow-covered roadways, interpret temporary construction signals and handle other tricky situations that many drivers encounter. (Google 2012)

Furthermore, in whatever ways Google guides these cars (e.g., using existing detailed mapping of roadways, rapid 360° surveillance using lidar (light-based radar), microwave communication between the guidance systems of nearby cars and with traffic control devices) the guidance requires dedicated computational powers and databases that lie beyond the competency and perceptions of human drivers. Wang (2014) notes that such guidance systems can collect data at a rate "between 20 and 60 megabytes per second."

11.3.3 Courtroom Reasoning

We consider the position of a jury R, composed of six or twelve individuals, that is trying to reach a verdict (the inference) in a trial by jury. The empirical domain centers on a courtroom process involving a judge, defendants, plaintiffs or prosecutors, and presentation of evidence and testimony regarding certain actions that the defendants may or may not have taken. The jury R cannot directly observe either the motivations of these actors or the actions they have taken. Elements of $\mathcal{I}_{R,t_i,t_f}$ generated prior to the start of testimony at t_i include such information as the jurors' own life experiences in judging the reliability and veracity of other people that can be applied to witnesses, plaintiffs, prosecutors, and defendants, and the jurors' life experiences that will help them later judge the likelihoods of scenarios. The jury reaches a judgment at time t_f. Information collected between t_i and t_f includes: trial testimony contents that are carefully monitored by the judge to be *admissible evidence* (e.g., admitted exhibits, statements of attorneys, witnesses) and recorded in a written transcript, portions of which may be read to the jury if their memories need to be refreshed. Individually, and later together, they add to $\mathcal{I}_{R,t_i,t_f}$ their thoughts about the reliability of witnesses based on such factors as demeanor, biases, and qualifications. An individual juror factors in, and shares, his or her sympathy or lack of sympathy with the prosecutor/plaintiff or defendant. In this instance, the consequences of a jury decision are borne largely by the defendant, plaintiff, or prosecutor, although individual jurors may feel their decisions weighing on their consciences. The information $\mathcal{I}_{R,t_i,t_f}$ available to the jury exceeds that available to individual jurors. The jury R is asked to refine its joint

judgment only to the non-quantitative extent of: *beyond a reasonable doubt* (criminal cases); *preponderance of the evidence* (civil suits); *clear and convincing* (e.g., when assessing the probability of the truthfulness of particular facts alleged during a civil lawsuit, such as a patent defense).

Controversial attempts have been made to render quantitative the process of jury judgment by introducing numerical probabilities (e.g., for pro number see Mode 1963, Finkelstein and Fairley 1971 and for anti number see Tillers 2011 and Tribe 1971). A 2011 book review by Balding 2011 noted that

> In the UK at least, recent judgments make clear that the courts remain hostile to numerical assessments of evidence based on any explicit use of subjective modeling assumptions.... (p. 358)

The approach set out in Section 11.5 better represents the assessments of a jury than would a quantitative probability assessment.

11.3.4 Medical Decision Making

The empirical domain focuses on the bodily state of a particular patient who presents at t_i, available diagnostic methods, and how the patient may react to any of a large set of possible interventions. An individual physician R has a $\mathcal{I}_{R,t_i,t_f}$ containing information accumulated prior to t_i from remembered textbook and laboratory knowledge of human anatomy and function, and mentored clinical experience with patients as a resident that is augmented through clinical experience gained through years of practice. $\mathcal{I}_{R,t_i,t_f}$ may contain selections from additional training in a specialization. $\mathcal{I}_{R,t_i,t_f}$ is then augmented by information collected in the time interval between t_i and t_f that includes laboratory and imaging results for the specific patient being treated, past medical records (how far back should the physician look?), the demeanor of the patient, and what s/he has to say about his or her condition. The physician R works with a complex array of different kinds of information, some specific to her/his patient, some derived from a large population of other patients, some quantitative, and some recalled from qualitative past clinical experience. As in the courtroom narrative, the consequences of the physician's decisions are largely borne by their patient and not themselves.

Evidence-based medicine (EBM)[7] attempts to quantify evidence gleaned from trusted clinical studies and combine it with physician experience to reach treatment decisions (see Iglehart 2005 for articles on this process and its limitations). In the *Bayesian threshold approach to clinical decision making* there are two thresholds, called "testing" and "test treatment", that are calculated from a physician's estimate of the relevant numerical probabilities of the possible diseases, the consequences of various treatments, and the risks (e.g., consequences of delayed or incorrect treatments and even of correct treatment).

[7] EBM has been defined (Greenhalgh 2010) as "the use of mathematical estimates of the risk of benefit and harm, derived from high-quality research on population samples, to inform clinical decision-making in the diagnosis, investigation or management of individual patients".

> Treatment should be withheld if the probability of disease is smaller than the testing threshold, and treatment should be given without further testing if the probability of disease is greater than the test-treatment threshold. A test should be performed (with treatment depending on the test outcome) only if the probability of disease is between the two thresholds. (Pauker and Kassirer 1980: p. 1109)

This approach formally recognizes that $\mathcal{I}_{R,t_i,t_f}$ is open to augmentation through additional diagnostic tests or trials with medications. However, the formal statistical threshold approach can collapse under the mental burden of its heavy demand for subjective quantitative probabilities. An example of this approach, reported by Gorry et al. (1973), for a study of a case of acute renal failure, included 1036 probabilities elicited from an "experienced nephrologist" that were added to $\mathcal{I}_{R,t_i,t_f}$!

> ....a clinician who performs well in diagnosis must have a good grasp of the relevant probabilities through his experience and familiarity with the literature....
>
> We have relied on subjective probabilities in this study because examination of the literature concerning acute renal failure has indicated that detailed quantitative information is, in a high percentage of instances, not yet available. (Gorry et al. 1973: p. 477)

Their recourse to subjective probability was a retreat from a desired frequentist probability. In addition, the authors needed to let their computer software normalize the elicited probabilities to unity as the elicited probabilities did not sum to one. We can have very little confidence in such a detailed elicitation of a huge number of probabilities.

In support of our conclusion, I turn to the first part of the preceding quotation about good performance in diagnosis implying a grasp of the probabilities. Cahan et al. (2005) studied the relationship between surely relevant clinical experience and an ability to produce explicit probability assessments when requested.

> Our objectives were to determine whether, when assessing *P* based on identical information (a written case scenario describing the history, physical examination, and ECG), differences would be found between senior and junior doctor populations in mean estimated P and variance. (Cahan et al. 2005: p. 1212)
>
> This study, based on case scenarios, did not find that medical expertise improved agreement among doctors when estimating the probability of disease in patients despite the common belief that senior physicians should have smaller interobserver differences in probability estimates. The wide variation observed calls into question the applicability of the threshold approach. (Cahan et al. 2005: p. 1211)

The disconnect between medical expertise and an ability to agree upon numerical probabilities of outcomes challenges the relevance of the latter. Sections 11.7, 11.8 describe alternative approachs in terms of upper and lower probabilities that replace precise individual probabilities and allow for incorporation of indeterminacy via the gap between the upper and lower probability for the same event.

11.3.5 Climate Forecasting

We all make frequent use of such vague expressions as "likely", "highly likely", "once in a blue moon", "very unlikely", "rarely", *probable*, "as likely as not", and *possible* in communication

about such matters as our intentions, medical diagnoses and efficacy of treatments, predictions about socioeconomic trends, and movements in financial parameters. There have been several empirical studies of such usages, including Brun and Teigen (1988), Budescu and Wallsten (1995), Reagan et al. (1989), as well as analytical approaches to characterizing such terms by Walley and Fine (1979), and Yalcin (2010).

The "Intergovernmental Panel on Climate Change (IPCC)" used such vague terms in their scientific evaluations of the evolution of global climate. The IPCC is convened to prepare "comprehensive assessment reports about the state of scientific, technical and socioeconomic knowledge on climate change, its causes, potential impacts and response strategies" (see plans for a fifth assessment concluding in late 2014; IPCC 2014). The "Guidance Notes for Lead authors of the IPCC Fourth Assessment Report on Addressing Uncertainties" (IPCC 2005), contain a Table 4 that is titled "Likelihood Scale". This table pairs "terminology" with "likelihood of the occurrence/outcome" as follows:

> Virtually certain $> 99\%$, Very likely $> 90\%$,
>
> Likely $> 66\%$, About as likely as not 33% to 66%,
>
> Unlikely $< 33\%$, Very unlikely $< 10\%$,
>
> Exceptionally unlikely, $< 1\%$.[8]

Formal aspects of the use of such comparative likelihood terms are examined in Sections 11.5 and 11.6.

The IPCC is empaneled at time t_i and completes its work at time t_f with a written synthesis report. Elements of $\mathcal{I}_{R,t_i,t_f}$ that precede t_i include dynamical meteorological theory, massive amounts of historical and current weather data, information about the factors (e.g., CO_2 and methane levels, surface temperature distribution, oceanic currents, industrial output, population and transportation) influencing the future evolution of the Earth's large-scale climate, and extensive reliance upon computer simulations. In the period between t_i and t_f, panelists make judgments of reliability and do not include in $\mathcal{I}_{R,t_i,t_f}$ all published theory and data. $\mathcal{I}_{R,t_i,t_f}$ involves judgments in extracting a small world of evidence from a much larger world by ignoring and simplifying many things, and negotiation among the panelists on these matters. This, of course, leaves these judgments open to later criticism and subsequent revision. The consequences of the judgments of the IPCC panelists are shared, as in the case of a jury. In this instance the consequences include the contrary opinions of colleagues, the reception by the public to whom their report is addressed, and their own sense of scientific and social responsibility.

The IPCC is commended for being open about the limitations to their ability to predict the future of our climate and for not hiding these limitations in the usual armored cloak of number. Their use of qualitative terms, explicated in terms of probability intervals, reflects their position that they cannot give scientific meaning to more precise specifications of probability. It is not that the IPCC had standard probabilities in mind and then converted them into linguistic terms for easier comprehension by the lay public. While people

[8] I have a drug insert on prescribing information that reports results of a trial with about 10,000 patients. The insert defines "infrequently" as between 1% and .1% and "rarely" as less than .1%, at variance with the IPCC definition of "exceptionally unlikely". However, I am comfortable with conversions of probability intervals into linguistic terms that depend upon the application context.

generally differ on how they would convert linguistic terms into probability intervals (e.g., see Budescu et al. 2009), IPCC has done a service by not obscuring the indeterminate aspects of their conclusions.

11.4 Grounding the Need for Alternative Theories of Probability

11.4.1 Precision and Accuracy

> At the margins of precision, the universe wavers.
>
> (Black 1970)

Precision refers to the fineness of mathematical distinctions offered by a measurement scale; e.g., a car speedometer dial is marked in increments of 5mph and readable to half that. *Accuracy* describes the faithfulness of the relationship between the measurement and the property being measured; e.g., the mass-produced speedometer may have been designed to read to within 2mph of the true road speed, with new, properly inflated tires of recommended size. However, in practice, speedometers are then adjusted to read higher[9] than the true speed so as to ensure they are not responsible for drivers exceeding posted speed limits. The end result is that the precision of car speedometers exceeds their accuracy.

Precision often exceeds accuracy; numerical IQ test results (with normal score of 100) have a precision of one part in a hundred, but surely their accuracy is far less for whatever concept of intelligence the IQ test is meant to measure. Engineering students are warned not to be led astray by computer-based computations that automatically report results with a precision of sixteen decimal digits even when based on input data accurate to three significant digits.[10]

An illustration of the difficulty of determining the proper precision for a probability statement or measurement is provided by Maher (2006: p. 146). Maher asks you to provide your supposedly numerical degree of belief about whether there is intelligent life in the Andromeda galaxy. The precision of your report is to be limited to the closest multiple of .001. He asserts, and I agree, that your degree of belief is higher than 0 and lower than 1 and that there is no $0 \leq n \leq 999$ such that your degree of belief is known to you to be higher than $.001n$ and lower than $.001(n+1)$. This example is a variation on the 'paradox of the heap' (*sorites* paradox, also called the 'little by little' argument).

Mathematical models or representations of chance and uncertainty, like that of standard probability, can be thought of as actual or potential measurements of the empirical phenomena at issue (e.g., the likelihood that a digital message received in noise can be decoded without error). We will say more about representation in Section 11.4.4.

[9] Higher by as much as 10% (Postkasse 2012) in Norway, while US standard allows up to 4% of maximum range, i.e., up to 5.6 mph on a 140 mph maximum.

[10] We omit an explanation of how the precision of the argument of a function being calculated transforms into a precision for the value of the function.

There is vast experience and much comfort with the use of real numbers[11] for empirical domains (e.g., the physical sciences) where both high accuracy and high precision have been achieved. However, expressing matters in terms of real numbers, that have a very high degree of precision, is often seriously misleading regarding the accuracy it automatically suggests is achievable for, say, many of the applications of probabilistic (or fuzzy) reasoning. Experimental physics accommodates to this by expressing measured physical quantities with a central numerical value and an error bar defined by two endpoint numbers that are derived from an assumption of Gaussian-distributed measurement errors of known variance. Psychology, in its use of a variety of measurement scales (e.g., see Luce 1996), offers a more sophisticated view of measurement than does physics and one that is more appropriate for probabilistic reasoning.

We start by examining precision and accuracy from frequentist and subjectivist (Bayesian) perspectives. These are the two most commonly invoked meanings for chance and uncertain phenomena. We conclude with the theory of measurement as developed largely in the social sciences and introduce the notions of *resolution* in the empirical and the mathematical domains.

11.4.2 Frequentist Perspective

Practicing frequentists[12] rely on real-valued probability, and their fundamental method for determining probabilities for a random experiment $\mathcal{E}$ starts with data generated by a sequence of n unlinked[13] repetitions of $\mathcal{E}$. For each event $A \in \mathcal{A}$, they calculate the *relative frequency* $r_n(A)$ as the ratio of the number of experiments having the outcome A to the total number n of repeated experiments. When estimating the probability of an event from a sequence of n repeated experiments, the *calculated relative frequency* measurement has a precision of $1/n$. However, the *statistical accuracy* is often estimated by some multiple of the standard deviation which is upper-bounded by $1/(2\sqrt{n})$. For large n, the experimentally determined precision is far finer than the accuracy of this measurement. Statistical practice often recommends using confidence interval estimates of an unknown probability rather than just a real-valued estimate. Nevertheless, it is fair to say that statisticians and probabilists act as if there is a true underlying real-valued probability.

More subtly, and perhaps more importantly, a frequentist probability estimate also has its accuracy limited by the potential instability, or temporal inhomogeneity, of the resulting relative frequencies, especially in the very long run. As noted by John Venn, a pioneer of the frequentist interpretation,

> ....that uniformity which is found in the long run.... though durable is not everlasting. Keep on watching it long enough, and it will be found almost invariably to fluctuate, and in time may prove as utterly irreducible to rule.... as the individual cases themselves. (Venn 1888: pp. 13, 14)

[11] A further extension of the precision of the reals to the *nonstandard reals*, that contain infinitesimals, has been used by Narens (2007) to represent comparative probability orderings (see Section 11.6).

[12] See La Caze's wide-ranging discussion of Frequentism in this volume.

[13] An intuitive term with formal probability counterparts of uncorrelated or independent.

This is a phenomenon that can be modeled by the methods to be discussed in Section 11.11.9. Very long-run failures of apparent convergence of relative frequencies are not rare. While real-valued probabilities are infinitely precise in principle, in statistical practice, efforts are made to estimate limits to the precision of a frequentist measurement so as to yield accurate probabilities at that precision. The idealization to real numbers has great formal mathematical convenience but is often a convenient fiction.[14] There is a need for judgment in frequentist modeling, but what is achievable falls far short of what is demanded by the subjectivist/Bayesian position discussed next.

11.4.3 Subjectivist Perspective

We hold the *subjectivist* meaning of probability[15] to refer to all mathematical characterizations of the mental state of uncertainty of an individual regarding a choice or decision. We reserve the *Bayesian* meaning for the special case of subjectivist probability where the mathematical characterization is provided by the unit interval $[0, 1]$ subset of the real numbers, the axioms of probability govern, and decisions are determined by *maximizing expected utility*. While Bayesians admit to the limited precision of the measurement process of *elicitation*, they are committed to using at least all the nonnegative rationals to represent *degrees of beliefs*[16] held by individuals—they view precision, and hence accuracy, as limitless. As evidence for this, I note the extensive Bayesian literature on *conglomerability*, an issue that only arises when conditioning on events drawn from an infinite partition of the sample space. Regarding the assumption of such numerical degrees of belief, recall Gertrude Stein's "there is no there, there" (in Everybody's Autobiography). Using the terminology of Levi (1985), I believe that the true issue is one of fundamentally limited accuracy or *indeterminacy* rather than one of imprecision.

What justifies the remarkable Bayesian position that real-valued probability can represent our degrees of belief? I suspect that the mathematical convenience of such an assumption provides its justification. Our mind is a consequence of our brain, nervous systems, and bodily chemical environment. It is informed by our traditional senses, but has additional inputs from other sources in the nervous system and in our body's chemical environment. All of our traditional senses have limited precision: our nose is far less sensitive than that of a dog; eyeglasses, microscopes, and telescopes bear witness to the limits of resolution of our visual acuity; our hearing has limited bandwidth and loudness sensitivity and may need assistive devices. Drugs can alter our perceptions by changing the chemical environment of the brain. Nevertheless, Bayesians grant the mind enormous powers of discernment regarding degrees of belief. We say "enormous" because Bayesians insist on modeling or representing these degrees of belief with an exceedingly precise real-valued, finitely additive probability measure. Bayesian elicitation of degrees of belief of individuals draws out approximations, perhaps in terms of intervals, to something for which there may be no

[14] Kolmogorov (e.g., see Li and Vitanyi 2008) recognized the need for a finite version of the asymptotic theory of frequentist probability. He provided one with the introduction of the computational complexity-based program and the notion of *Kolmogorov complexity*.

[15] See the extensive discussion of Subjectivism (17) in this volume.

[16] See Eriksson and Hâjek (2007) for a critical discussion of approaches to this concept.

faithful or accurate characterization in such terms. Elicitation can create that which it seeks to elicit—sufficient encouragement or intimidation can elicit almost any desired response.

What though of what is meant to be the decisive argument provided by the possible existence of *Dutch Books*? A Dutch Book is an offer of a combination of wagers, each of which you favor, but of which if you accept the combination, then you cannot profit. A theorem then proves that if you acted as if you had a common probability model for the various events determining payoffs, utilities (special numerical values) for these payoffs, and given a choice you preferred only those wagers maximizing expected utility, then you could not fall prey to a Dutch Book but that in the absence of such an approach you would be susceptible.[17] However, I am not "rationally" compelled by my choices of wagers I consider favorable to take an unfavorable combination of such wagers that is itself a wager I have yet to assess, unless that compulsion be provided externally to my beliefs. I can guard myself against such exploitation.

Bayesians have an imperial reach that sees subjective numerical probability as the answer to all questions. In this, their reach exceeds both their grasp and the more circumspect reach of the frequentists. The overreach of the Bayesian program, spanning from how we think to how we characterize the physical world, is evidenced in the following two quotations.

> We have outlined an approach to understanding cognition and its origins in terms of Bayesian inference.... (Tenenbaum et al. 2011: p. 1285)

> The physical law that prescribes quantum probabilities is indeed fundamental, but the reason is that it is a fundamental rule of inference—*a law of thought*—for Bayesian probabilities. (Caves et al. 2002: p. 5)

We should support the use of real-valued probability, as a mathematically convenient description of degrees of belief, only when there is clear recognition of the resulting discrepancy between calculated precision and actual accuracy.

11.4.4 Perspective from the Theory of Measurement

To represent as probability represents chance, is to measure, to make knowable in principle. We adopt the viewpoint of the *theory of measurement* as described in Section 1.2.2 of Krantz et al. (1971), the first of three monographs on measurement theory by permutations of the four authors of Krantz et al. (1971). The following are the components of this theory:

1. an *empirical domain* of real-world phenomena of interest and a set of significant relationships between them, as illustrated in Section 11.3;
2. a chosen *mathematical (co)domain*, e.g., a set of unordered pairs (A, B) for the concept of A and B are "related"; integers $\mathbb{N}$ for labeling distinct entities; nonnegative numbers $[0, \infty)$ for ages, lengths, weights, masses; the unit interval $[0, 1]$ for standard probability; an ordered collection of sets for "at least as probable as" as illustrated in Section 11.6.2;

[17] Hájek (2008) offers an analysis of the limitations of such arguments.

3. a correspondence from the empirical domain to the mathematical codomain that maps[18] those relationships of interest in the empirical domain into particular relationships between elements of the mathematical domain (e.g., using real numbers in $[0, 1]$ can mean that the probability of A has infinite precision in that it might be a transcendental or incomputable decimal);
4. and this correspondence attempts to minimize mathematical identification of spurious empirical relationships.

The empirical domain knowledge assembled by Reasoner R attributes properties to things, and to relations between them, only up to a certain fineness of grain. The Reasoner's corresponding choice of mathematical model must be flexible enough to reflect that fineness of grain. If the model cannot express fine enough distinctions then it will hide from R distinctions that may prove to be important. If the model enables excessively fine distinctions then it suggests empirical domain distinctions that are purely artifacts of the model, are empirically meaningless, and can readily mislead R into pursuing these phantom distinctions. For instance, beyond the test score itself, is there any useful sense in which an individual tested to have an IQ of 128 is "more intelligent than" one with an IQ of 127?

11.4.5 Resolution in the Empirical and Mathematical Domains

We introduce two concepts of *resolution*,[19] an informal one for the empirical domain and a formal one for the mathematical domain.

11.4.5.1 *Empirical Resolution*

The qualitative concept of the *empirical resolution* of a property of, or relation holding between, elements of the empirical domain refers to a measure of your limited ability to discern, distinguish, discriminate, or differentiate on the basis of that property or relation. The term *empirical resolution* is intended as a neutral cognate (because of its lack of use) of such more familiar terms as: indeterminacy (used by Levi 1985), imprecision (used by Walley 1991), fuzziness,[20] and ill-definedness. We cannot offer a formal approach to empirical resolution that accounts for the highly varied empirical domains that are

[18] Mathematically, this mapping is a *homomorphism*, a many-to-one mapping that preserves relationships of interest in the empirical domain; e.g., on the Bayesian account a probability measure μ must reflect greater strength of belief through larger numerical values and also preserve coherent preferences for wagers when used to calculate expected utility. Uncertainties about diverse matters may yet be represented by the same number.

[19] This choice of word is motivated by the optical concept of spatial resolution of a lens as the shortest distance between two distinct points in the field of vision that can still be seen through the lens as distinct entities.

[20] Zadeh introduced the term "fuzzy". He characterized fuzzy entities by degrees of membership specified by real-valued functions that make their context-dependent meaning exceedingly precise, in contradiction to the low accuracy intended when we use such a vague term.

not themselves formal systems.[21] As a mundane example, consider an empirical domain of measuring or representing "human lifespan". Human lifespan has limited numerical resolution for persons who have died and even less resolution for those still alive. How accurately do we know when a person's life started, on whatever notion we have of this onset, and how accurately can we identify the time of death? Neither endpoint is identifiable to within, say, a second, or even much longer, for almost all of those who have already died. The accuracy or precision apparently inherent in any real-number statement of lifespan is spurious.

Turning to the empirical domain concepts of "likelihood of occurrence", "degree of belief", and " epistemic support", I assert that they also have limited resolution. An assessment of the likelihood of the existence of extraterrestrial intelligent life (e.g., see Sagan 1973 and Maher 2006 on Andromeda), whether objective or subjective, has less empirical resolution than does the assessment of the likelihood of "red" on a specific roulette wheel, at a reputedly honest casino, that may have become slightly unbalanced after long usage. Limited empirical resolution can be due to: the finiteness of frequentist data; possible unstable long-term conditions such as unknown drift and cyclical variation; intrinsic vagueness and temporal instability of personal belief; limited and contradictory assertions of propositions in $\mathcal{I}_{R,t_i,t_f}$ that do not lead to fully trustworthy inferences, let alone precise deductions; absence of common agreement on the meaning and expression of inductive inference in situations of inconsistent $\mathcal{I}_{R,t_i,t_f}$ and goals that may be understood only vaguely. In almost all empirical situations (even one as seemingly straightforward as the weight of a person at a given time), when carefully examined, there is a large variety of overlooked, confounding, and limiting phenomena (e.g., gravitational force variations with location, vibration at the scale, the person's inhaling, exhaling, or transpiration) that interact with the property or relationship of interest, and that we ignore (usually unthinkingly). These phenomena limit the accuracy of what is meaningful regarding empirical resolution, and, thence, what is meaningful regarding the precision of a mathematical representation.

11.4.5.2 *Mathematical Resolution*

The *mathematical alternatives to standard probability* that are our focus provide the flexibility to express cognizance of the degrees of empirical domain resolution that are attuned to particular manifestations of such concepts as chance, uncertainty, partial inference, propensity, vagueness, and indeterminacy. The mathematical codomains of standard probability (e.g., $[0,1]$, a σ-algebra of subsets $\mathcal{A}$, and a probability measure μ defined by the axioms of standard probability) are often unable to match the resolution of the above-mentioned empirical domain probabilistic concepts. It is inconsistent to match a qualitative degree of empirical resolution to a quantitative degree of *mathematical resolution*. A quantitative mathematical model offers fine distinctions that may have no correspondence in the empirical domain and that then mislead us as to what can be defensibly or meaningfully asserted about the empirical domain.

[21] Those empirical domains addressed by logical or epistemic probability might provide exceptions.

11.5 Unconditional Classificatory or Modal Probability

11.5.1 Motivation and Overview of *Classificatory* or *Modal* Probability

The terms *possible*, *probable*, and *improbable* are examples of *classificatory* or *modal* probability. The approach taken in Walley and Fine (1979) is closest to this presentation. An early paper on this subject is Hamblin (1959) and a linguistically-oriented perspective is found in Black (1967). Modal probability is also discussed from the viewpoint of betting behavior in Section 4.4 of Walley (1991), and betting behavior need not be subjectively based. I axiomatize (bring into the mathematical codomain) three intentionally vague terms describing a range of uncertainties for some events in the empirical domain (e.g., "it will rain here by noon tomorrow", "X spoke truthfully"). I do not follow the extensive "fuzzy" literature that defines "possible" and "probable" in terms of real-valued functions—a conversion of the intrinsically vague into the most highly detailed descriptions that have no warrant beyond familiarity and wishful thinking. Nor do I follow the route of modal logics and possible worlds interpretations (e.g., see Yalcin 2010).

11.5.2 Axioms for *Possible* and *Probable*

Subsets of the following six axioms, describing generally agreed-upon properties of *possible* and *probable*, will be used to provide at least a partial formal characterization of *possible* and *probable*.

I. Impossible Event The empty set $\emptyset$ is neither possible nor probable.

II. Certain Event The sample space Ω is both possible and probable.

III. Subsets of Possible Sets $A = B \cup C$ is possible if and only if at least one of B or C is possible.

Corollary 1 (Possibility Agrees with Set Inclusion). *If B is possible and $A \supseteq B$, then A is possible.*

Proof. Since $A \supseteq B$ implies $A = B \cup (A - B)$, the corollary follows immediately from the "if" part of Axiom III.

Lemma 1 (Either A or A^c are Possible). *For every subset A of Ω at least one of A or A^c is possible.*

Proof. By Axiom II, Ω is possible. For any A, $\Omega = A \cup A^c$. Hence, by the "only if" part of Axiom III, at least one of A or A^c is possible.

It does not make sense to have Axiom III be a property of 'probable'. Having A be 'probable' does not imply that any partition of $A = B \cup C$ into smaller sets will yield a B or C that is 'probable'.

IV. Constraint on Probable For every set A, at most one of A and A^c is probable.

(Thus, *probable* is 'more probable than not'.)

V. Supersets of Probable Sets If A is probable and $B \supseteq A$, then B is probable.

VI. Coordination of Probable with Possible If A is probable then A is possible.

In what follows, we will have reason to distinguish between Ω finite, countably infinite, and uncountably infinite.

11.5.3 Representing Possible

Axioms I and II are essential to any meaning of a "possible set/event"; an event that necessarily cannot occur cannot be possible, while an event that necessarily must occur must be possible. Axiom III just asserts that if B is possible, for whatever reason, and $A \supseteq B$ then A is possible since A occurs whenever the possible event B occurs.

Definition 1 (Possible Event Collection). *A collection $\mathcal{P}$ of subsets of Ω is a collection of* possible sets/events *if and only if it satisfies Axioms I, II, III.*

Theorem 1 (Representing *Possible*). *(a) For arbitrary nonempty Ω and nonempty subset W, the collection*

$$\mathcal{P} = \{A : (\exists w \in W)\ w \in A\}$$

satisfies Axioms I, II, III and therefore is a collection of possible sets.

(b) If Ω is finite, then a collection $\mathcal{P}$ of subsets is a collection of possible sets only if there exists a nonempty $W \subset \Omega$ and $\mathcal{P}$ is defined as in (a).

Proof. (a) To prove (a) assume a given nonempty $W \subseteq \Omega$ and consider the collection $\mathcal{P} = \{A : (\exists w \in W)\ w \in A\}$. Clearly, $\emptyset \notin \mathcal{P}$ and $\Omega \supseteq W$ and is in $\mathcal{P}$. Hence, Axioms I, II are satisfied for $\mathcal{P}$. By construction of $\mathcal{P}$ if $B \cup C \in \mathcal{P}$ then there is at least one $w \in W$ that is also in $B \cup C$. Hence, w must be in at least one of B, C, and at least one of B, C is then in $\mathcal{P}$, satisfying the 'only if' part of Axiom III. For the 'if' part of Axiom III, if B is in $\mathcal{P}$, then there exists $w \in B \cap W$. It is immediate that for any C, $B \cup C$ also contains w and is in $\mathcal{P}$. Thus, Axiom III is satisfied, and $\mathcal{P}$, as constructed, is a set of *possible* sets.

(b) To prove the necessity of the construction of a collection of possible sets in (a) for finite Ω, assume that $\mathcal{P}$ is a given set of *possible* subsets of Ω satisfying the three defining axioms. Note that by Corollary 1, any superset of a possible set is also possible. Without loss of generality, assume that Ω has $n \geq 1$ elements. Construct W as the set of all of the individual points (sets of cardinality 1) in Ω that are in the given $\mathcal{P}$. To prove that W must be nonempty, take any $n \geq k \geq 1$ for which there is an $A \in \mathcal{P}$ and A has k points (cardinality $\|A\| = k$). There must be at least one such set A since $\Omega \in \mathcal{P}$ by Axiom II. Enumerate the elements of $A = \{\omega_1, \ldots, \omega_k\}$. For any $1 \leq j \leq k$, by Axiom III it is the case that at least one of

ω_j or $A - \{\omega_j\}$ is in $\mathcal{P}$. If it is ω_j then we have shown that W is nonempty. If it is not ω_j then $A - \{\omega_j\} \in \mathcal{P}$ and we can iterate by extracting one of the remaining elements of $A - \{\omega_j\}$. This process terminates in no more than $k \leq n$ steps with the identification of an element of W, proving that W is nonempty, and also that every $A \in \mathcal{P}$ has a nonempty intersection with W.

□

If $W = \Omega$, then A is possible if and only if it is nonempty. At the other extreme, if W has only a single element w^* then A is possible if and only if it contains w^*.

Removing the restriction to finite Ω permits instances of possible set collections that are not generated by any W as in Theorem 1(a). Consider the collection $\mathcal{Q}$ of all infinite subsets of an infinite Ω. We claim that $\mathcal{Q}$ satisfies Axioms I, II, III. As $\emptyset$ is not infinite and Ω is infinite, Axioms I, II is satisfied. If $A \cup B$ is infinite then at least one of A or B are infinite and therefore possible. If A is infinite and B is arbitrary, then $A \cup B$ is infinite and therefore possible. Thus, Axiom III is satisfied and $\mathcal{Q}$ is a collection of possible sets. However, there does not exist a W as used in Theorem 1. Any such putative W must have a nonempty intersection with all infinite subsets of Ω. As this W is perforce nonempty it would also render all of its finite subsets possible, in contradiction to $\mathcal{Q}$ not containing any finite sets.

11.5.4 Relation of *Possible* to Probability

$\mathcal{P}$ being a collection of *possible* subsets of countable Ω is ensured if there exists a probability measure μ, defined for all subsets of Ω, such that $A \in \mathcal{P}$ if and only if $\mu(A) > 0$. As $\mu(\emptyset) = 0, \mu(\Omega) = 1$, this definition satisfies Axioms I, II. Satisfaction of Axiom III follows from the fundamental facts

$$\mu(A) + \mu(B) \geq \mu(A \cup B) \geq \max(\mu(A), \mu(B)).$$

Thus, $A \cup B \in \mathcal{P}$ implies that $\mu(A \cup B) > 0$, which in turn, implies that at least one of $\mu(A), \mu(B)$ is positive. Conversely, having at least one of A, B be possible forces $\mu(A \cup B) > 0$ and $A \cup B$ is possible.

If Ω is countable and $\mathcal{P}$ can be described in terms of a subset W, then from Theorem 1(a) we can alternatively describe $\mathcal{P}$ by choosing any μ defined by its probability mass function

$$p : \Omega \to [0, 1], \ \sum_{\omega \in \Omega} p(\omega) = 1, \ (\forall \omega \in W) p(\omega) > 0, \ \sum_{\omega \in W} p(\omega) = 1,$$
$$\mu(A) = \sum_{\omega \in A} p(\omega).$$

The disadvantage of this representation is that the choice of μ is quite arbitrary.

Interestingly, there exist $\mathcal{P}$ on (at least countably) infinite Ω, satisfying the axioms for a set of possible events that cannot be represented by any standard probability measure μ. To verify this, use the previous example of A being possible if and only if it is an infinite set. It was found above that this construction satisfies Axioms I, II, III. A probability representation by μ (that must be defined for all subsets of Ω) would require that for all

finite B, $\mu(B) = 0$. There does not exist such a (countably additive) standard probability μ. Hence, possibility cannot always be modelled in terms of standard probability.

11.5.5 Representing *Probable*

Definition 2 (Probable Events). *(a) A collection* Π *of subsets of* Ω *is a collection of* probable events *if and only if* Π *satisfies Axioms I, II, IV, V.*

(b) A nonempty collection $\mathcal{W}$ *of subsets of* Ω *is a* probable-generator set *if and only if:*

(i) $\mathcal{W}$ *contains a single nonempty set* W*, or*

(ii) every pair of sets W_1, W_2 *in* $\mathcal{W}$ *has a nonempty intersection,*

(iii) and, in either case, for any $W \in \mathcal{W}$ *no proper subset of* W *is in* $\mathcal{W}$*.*

An example of a probable-generator set is (omitting internal set braces) $\mathcal{W} = \{01, 02, 12\}$ for $\Omega = \{0, 1, 2, 3\}$.

Theorem 2 (Representing *Probable*). *(a) For* Π *to be a collection of the probable subsets of an arbitrary nonempty* Ω*, it suffices that there exist a* probable-generator set $\mathcal{W}$ *such that*

$$\Pi = \{A : (\exists W \in \mathcal{W})\ A \supseteq W\}.$$

(b) If Ω *is a finite set, then the condition (a) is also necessary.*

The set Π of probable sets generated by the example above of $\mathcal{W}$ is

$$\Pi = \{01, 012, 013, 02, 023, 12, 123, 0123\}.$$

Proof. (a) It is easy to show "sufficiency" in that a generator set $\mathcal{W}$ will induce a collection Π of *probable* sets satisfying Axioms I, II, IV, V. Axioms I, II follow immediately from the definition of Π. Observe from the postulated overlap between any two sets in $\mathcal{W}$ that it follows that any two probable sets A, B induced by $\mathcal{W}$ will have a nonempty intersection. Hence, Axiom IV is satisfied. If $A \subset B$ and A is probable, then it contains a set, say, $W \in \mathcal{W}$, and W is also a subset of B. Therefore, B is probable, and Axiom V is satisfied.

(b) To prove "necessity" for the given representation in the case of finite Ω, we need to show that given any Π for finite Ω, we can construct $\mathcal{W}$. Given Π satisfying Axioms I, II, IV, V, for a finite Ω, we iteratively construct $\mathcal{W}$ that will be a probable-generator set. For the initial step $\mathcal{W}_1$, choose the probable sets of cardinality 1 in Π, if there are any. At step k, include in $\mathcal{W}_k$ the probable sets of cardinality k that do not have proper subsets that are already contained in $\cup_{i<k}\mathcal{W}_i$. This process terminates at iteration $\|\Omega\|$, the finite size of Ω. Define

$$\mathcal{W} = \bigcup_{k=1}^{\|\Omega\|} \mathcal{W}_k.$$

The resulting $\mathcal{W}$ is a collection of subsets of Π such that every set $A \in \Pi$ contains a set in $\mathcal{W}$. Furthermore, if $B \notin \Pi$ then it cannot contain any set in $\mathcal{W}$. Assuming that there were such a B containing a W, then by our construction and Axiom V, we reach the contradiction that

it would have to be in Π. Finally, the collection $\mathcal{W}$ is minimal in that removing any element w from any $W \in \mathcal{W}$ yields $W - w \notin \mathcal{W}$, as required by Definition 2b(iii). □

11.5.6 Relation of *Probable* to Probability

Π being a collection of probable sets is ensured if there exists a probability measure μ *on all subsets* of a countable Ω and a threshold $1/2 < \lambda \leq 1$, such that $A \in \Pi$ if and only if $\mu(A) \geq \lambda$. It is elementary to verify that this construction satisfies Axioms I, II, IV, V that define *probable* sets. If choosing $\lambda \leq 1/2$ induced a different collection Π than could be induced by a λ minimally greater than $1/2$, then the choice less than $1/2$ would allow for a probable A and A^c, in contradiction to Axiom IV.

A simple counterexample of Walley's (see Walley and Fine 1979), to the ability of standard probability to describe all finite collections of probable events, is provided by the following example of *probable*:

$$\Omega = \{0, 1, 2, 3, 4, 5, 6\};$$
$$\mathcal{W} = \{023, 134, 245, 356, 460, 501, 612\}.$$

Define as *probable* exactly those sets containing at least one of the seven sets in $\mathcal{W}$. By Theorem 2, $\mathcal{W}$ defines a collection Π of *probable* sets.

To show that there cannot be a probability measure representation of *probable* as we have just defined it, observe that each of the seven elementary outcomes in Ω occurs exactly three times in the collection $\mathcal{W}$. Hence, no matter what probability measure μ you select

$$\sum_{W \in \mathcal{W}} \mu(W) = 3.$$

Every probability representation of *probable* must imply $\mu(A) > 1/2$ for A to be probable. This provides a lower bound to any such representation that yields

$$\sum_{W \in \mathcal{W}} \mu(W) > 7(\frac{1}{2}) > 3 = \sum_{W \in \mathcal{W}} \mu(W),$$

and the desired contradiction is reached.[22]

11.5.7 Consistency of *Probable* with *Possible*

We would expect that if A is probable then it is also possible, although not conversely. This is what is asserted in Axiom VI of Section 11.5.2. However, Axiom VI need not hold if *probable* and *possible* are defined without explicit regard to each other. We start with a definition of *probable* and then show that we can define *possible*, in many ways, such that all probable events are also possible.

[22] A smaller example containing only four sets was constructed by Joel Seifras that yields a less forceful contradiction, in which $\sum_{W \in \mathcal{W}} \mu(W)$ evaluated one way has the value 2 and evaluated another way has a value strictly greater than 2.

Theorem 3. *When the* probable sets *in* Π *satisfying Axioms I, II, IV, V are generated by some* $\mathcal{W}$*, then they are also* possible sets *satisfying Axioms I, II, III if we choose a set* W*, defining possible sets via Theorem 1(a), so that it has a nonempty intersection with every set in* $\mathcal{W}$ *generating* Π.

Proof. Choosing W, defining *possible*, to have a nonempty intersection condition with each $W' \in \mathcal{W}$, defines each such W' as possible. Hence, each set containing a W' (a *probable* set) is also *possible* as it contains an element of W. The collection of possible sets will generally be larger than the original collection of probable sets. □

Conversely, we can create probable sets from possible sets when the latter have the representation given in Theorem 1 of Section 11.5.3. Simply take $\mathcal{W}$, defining "probable", to contain the single set W, defining "possible". This will render some of the possible sets as probable.

11.5.8 Representing *Highly Probable*

We expect that a highly probable event A should also be probable. We tentatively make sense of the modality of *highly probable* by assuming that it is a restriction of a pre-specified modality of *probable*. Given Ω, assume that $\mathcal{W} = \{W\}$ is a probable-generator set as defined in Definition 2(b) of Section 11.5.5. The weak definition we propose next for *highly probable* renders any particular representation of "probable" at least slightly more stringently.

Definition 3 (Highly Probable)**.** *Given* $\mathcal{W} = \{W\}$ *defining* probable *on subsets of* Ω*, define a* highly probable-generator set $\mathcal{W}^* = \{W^*\}$ *satisfying*

$$(\forall W^* \in \mathcal{W}^*)(\exists W \in \mathcal{W})\, W^* \supseteq W \text{ and } (\exists W^* \in \mathcal{W}^*)(\forall W \in \mathcal{W})\, W^* \supsetneq W.$$

Define the collection Π^* *of* highly probable *sets through*

$$\Pi^* = \{A : (\exists W^* \in \mathcal{W}^*)\, A \supseteq W^*\}.$$

Note that any two sets in $\mathcal{W}^*$ have a nonempty intersection because they each contain sets drawn from $\mathcal{W}$ that have a nonempty intersection. Hence, $\mathcal{W}^*$ satisfies the conditions for being a probable-generator set. This construction of *highly probable* sets strengthens a given condition of being *probable* while ensuring that Axioms I, II, IV, V are still satisfied.

11.5.9 Mathematical Resolution in Modal Probability

The concepts axiomatized above have a *mathematical structure* having the least resolution of any of the remaining probability-like alternatives to standard probability that we will discuss. When $\mathcal{P}, \Pi, \Pi^*$ are definable through a probability measure and threshold on a finite Ω, they are also definable through infinitely many other probability measures and threshold values. However, for $\|\Omega\| = n$ there are only finitely many ways to select a collection of subsets of Ω that satisfies the axioms of either *possible*[23], *improbable*,

[23] In this case there are $2^n - 1$ ways to select a nonempty W.

probable, or *highly probable*. Hence, the infinite number of choices required to specify a specific probability measure μ has been reduced to a finite number—a great reduction in mathematical resolution. Furthermore, this low degree of mathematical resolution agrees well with the low degree of empirical resolution in assertions of *improbable* or *probable* that is found in such studies of variable interpersonal usage as Budescu and Wallsten (1995). If we remain committed to standard probability, then we would either have to commit to a choice of a specific probability measure and threshold, whose specificity has no warrant in the little information being communicated by the modal concept, or to awkwardly characterize these modal concepts by sets of standard probability measures.

11.6 Unconditional Comparative Probability (CP)

11.6.1 Motivation and Overview of CP

Comparative probability (CP) is a binary relation of "at least as probable as" holding between two events. Introductions to CP can be found in Kraft et al. (1959), chapter II of Fine (1973), Fine (1977), Kaplan and Fine (1977), Walley and Fine (1979), and Kumar (1982). A more recent summary is that of Regoli (2000).

Keynes (1921), as early as 1919, explored conditional comparative probability for hypothesis h and evidence e propositions, with $h_1|e_i \succsim h_2|e_2$ read as "hypothesis h_1, given evidence e_1, is at least as probable as hypothesis h_2 given evidence e_2. Keynes' CP carried an *epistemic* interpretation. James Hawthorne, in his contribution to this volume, explores this line of conditional qualitative probability with an epistemic interpretation. Bruno de Finetti (1931) considered CP with a view to creating qualitative probability that would formalize a subjective interpretation. De Finetti hoped (knowing he had not proven it to be true) that CP would provide an intuitively sound subjective numerical probability representation, based upon a determination of which of two events is the more probable, that would be a more stable[24] representation of degrees of belief than is an assignment of numerical values to each of the two events.

11.6.2 Rationality Axioms for CP

A CP random experiment $\mathcal{E} = (\Omega, \mathcal{A}, \succsim)$ replaces a numerical probability description by a complete ordering of events to be axiomatized below. We use the symbol $\succsim$ to stand for a binary relation between sets in an algebra $\mathcal{A}$ of subsets of a sample space Ω. $A \succsim B$ is read as "A is at least as probable/likely as B". The binary relation $B \precsim A$ means the same as $A \succsim B$.

[24] More stable in the sense that the CP representation for the case of strict inequalities between distinct events, when it is compatible with a probability representation (see Section 11.6.4), is invariant to small perturbations in the probability representation.

"A is more probable than B" is denoted by $A \succ B$ and "A is as probable as B" is denoted by $A \sim B$.

$$A \succ B \iff A \succsim B \text{ and false } B \succsim A;$$
$$A \sim B \iff A \succsim B \text{ and } B \succsim A.$$

The following axioms for $\succsim$ are due to Bruno de Finetti (see de Finetti 1931).

CP1. (non-triviality) $\Omega \succ \emptyset$.

CP2. (improbability of impossibility) $(\forall A \in \mathcal{A}) A \succsim \emptyset$.

CP3. (reflexive) $(\forall A \in \mathcal{A}) A \succsim A$.

CP4. (comparability/completeness) $A \succsim B$ or $B \succsim A$.

CP5. (transitivity) $A \succsim B$ and $B \succsim C \implies A \succsim C$.

CP6. (cancellation) $A \succsim B \iff A - B = A \cap B^c \succsim B - A = B \cap A^c$.

A simple example based upon $\Omega = \{a, b, c\} = abc$ (letting, say, ab denote the set $\{a, b\}$) is

$$\emptyset \prec a \sim b \prec ab \prec c \prec ac \sim bc \prec abc = \Omega.$$

Axioms CP3,4,5 establish the total or complete order properties of $\succsim$. Axioms CP1,2,6, provide the additional properties of a comparative probability ordering, the key one being CP6. Together these axioms suffice when $\mathcal{A}$ is a finite set. Understanding the finite set case is the first task of CP.

Two implications of these axioms are that the CP ordering agrees with the partial ordering by set inclusion,[25]

$$A \supseteq B \implies A \succsim B,$$

and that CP is order-reversing under complementation,[26]

$$A \succsim B \iff B^c \succsim A^c.$$

This setup can be enriched by extensions to infinite sample spaces (e.g., see Narens 2007) and by the introduction of a definition for independent CP experiments (see Kaplan and Fine 1977). Surprisingly, irrespective of independence, it is not always possible to combine two CP descriptions of two random experiments so as to form a joint experiment (see Kaplan and Fine (1977)).

[25] By cancellation, $A \succsim B$ if and only if $A - B \succsim B - A$. When $A \supseteq B$, this last statement holds if and only if $A - B \succsim \emptyset$, and this is true by CP2.

[26] If you assume to the contrary that there exist A, B with $A \succsim B$ and $A^c \succ B^c$, then by cancellation, these two relations are equivalent to $AB^c \succsim A^c B$ and $A^c B \succ AB^c$, and the contradiction is reached.

11.6.3 Failure of Calibration of One CP Experiment in Terms of Another

Any two standard probability random experiments, $\mathcal{E}_1, \mathcal{E}_2$, for example, corresponding to random variables X_1, X_2, can always be combined into a single random experiment $\mathcal{E}$ for the pair $X = (X_1 \times X_2)$ in which there is independence between any event $((X_1 \in A_1) \times \Omega_2)$ and $(\Omega_1 \times (X_2 \in A_2))$. Surprisingly, however, there need not exist a joint experiment $(\succsim, \mathcal{E})$ combining two CP descriptions $\succsim_1, \succsim_2$ of two random experiments $\mathcal{E}_1, \mathcal{E}_2$, even when the sample spaces are finite and we allow for dependent combinations (Kaplan and Fine 1977). This observation is key because otherwise we could "numerically calibrate" any given CP experiment $\mathcal{E}_1$ by independently combining it with a CP experiment $\mathcal{E}_2$ having a uniform probability distribution as a representation. It is common in discussions of subjective probability to assume that such calibration is innocuous. That it is not an innocuous move should come as little surprise given the heavy lifting this assumption would allow you to do.

11.6.4 CP and Standard Probability

Definition 4 (Relation to Standard Probability Measures)**.**

1. *A standard probability measure* μ agrees *with (or* represents*) the CP order* $\succsim$ *between sets in an algebra* $\mathcal{A}$ *of subsets of* Ω *if*

$$(\forall A, B \in \mathcal{A})\ A \succsim B \iff \mu(A) \geq \mu(B).$$

   When this is the case we say that $\succsim$ *is* additive.
2. *If there exists* μ *such that*

$$(\forall A, B \in \mathcal{A})\ A \succsim B \implies \mu(A) \geq \mu(B),$$

   then $\succsim$ *is* almost additive.
3. *If a CP order* $\succsim$ *is neither additive nor almost additive then we say that it is* non-additive.

For example, it is immediate that for any μ agreeing with our previous (p. 225) CP ordering of *abc*, we must have

$$0 < \mu(a) = \mu(b), \mu(a) + \mu(b) < \mu(c), \text{ and always } \mu(a) + \mu(b) + \mu(c) = 1.$$

From this we conclude that $0 < 2\mu(a) < \mu(c), 2\mu(a) + \mu(c) = 1$, implying

$$0 < 2\mu(a) = 1 - \mu(c) < \mu(c),$$
$$\text{hence, } 1 > \mu(c) > 1/2 \text{ and } \mu(a) = \mu(b) = \frac{1}{2}(1 - \mu(c)).$$

Typically, as in this case, an agreeing probability measure is not unique.

11.6.5 When Is CP Additive?

While every probability measure μ represents a CP ordering of events, it is not the case that every CP ordering can be represented by a probability measure.

CP is always additive for a random experiment on a sample space Ω that has no more than four elements; e.g., say, one or two tosses of a coin. However, even for a sample space Ω as simple as one toss of a die (six possible outcomes), there exists a CP order $\succsim$, satisfying the CP axioms, for which no agreeing probability measure exists! This result was first established by Kraft et al. (1959), who provided a specific CP ordering we call KPS5 on a sample space of five points, that is non-additive as we have defined it, as well as a CP ordering we call KPS6 of six points, that is non-additive in an even stronger sense.

The smallest possible Ω that supports non-additive CP orders requires five elements. KPS5 is defined on $\Omega = abcde$ by

$$\begin{aligned}&\emptyset \prec a \prec b \prec c \prec ab \prec ac \prec d \prec ad \prec bc \prec e \prec abc \prec bd \prec cd \prec ae \prec abd\\ &\prec be \prec acd \prec ce \prec bcd \prec abe \prec ace \prec de \prec abcd \prec ade \prec bce \prec abce\\ &\prec bde \prec cde \prec abde \prec acde \prec bcde \prec abcde = \Omega.\end{aligned}$$

The property of order reversal under complementation enables us to determine through complementation the last half of the CP order starting with *acd* from the first half ending with *be*. To verify that KPS5 does not have any probability representation, consider the following four inequalities:

$$ac \prec d,\ ad \prec bc,\ cd \prec ae,\ be \prec acd.$$

If there exists a probability representation μ then it must follow that

$$\begin{aligned}&\mu(a)+\mu(c) < \mu(d),\ \mu(a)+\mu(d) < \mu(b)+\mu(c),\\ &\mu(c)+\mu(d) < \mu(a)+\mu(e),\ \mu(b)+\mu(e) < \mu(a)+\mu(c)+\mu(d).\end{aligned}$$

Adding left-hand sides L together and then right-hand sides R together yields

$$L = 2\mu(a)+\mu(b)+2\mu(c)+2\mu(d)+\mu(e) < R = 2\mu(a)+\mu(b)+2\mu(c)+2\mu(d)+\mu(e).$$

We find that $L < R$ and $L = R$, and the desired contradiction is reached.

Scott (1964), Theorem 4.1, provided necessary and sufficient conditions for a CP order to be additive and thereby rule out examples like KPS5,6. These necessary and sufficient geometric conditions (based upon hyperplane separation of multidimensional convex sets) generated a countably infinite system of linear inequalities that had to be satisfied for a CP order to be additive. In effect, this system amounted to moving from the equivalent of binary-valued events (they either occurred or they did not) to consistency conditions for integer-valued random variables. Restated, the condition for the existence of a representing probability requires consistency for all *expected values* of integer-valued random variables defined on Ω. The existence of a probability representation for a CP order representation question can be resolved only by requiring an infinite enlargement of the set of CP axioms that effectively move us beyond the realm of events to one dealing with comparing

expectations of random variables. Stated differently, there is no finite system of elementary axioms for CP that guarantee additivity and the existence of a representing probability measure.

11.7 Belief Functions and CP

11.7.1 Basics of Belief Functions

A general representation of CP through real-valued functions is available from *belief functions*, a special class of upper/lower probabilities (see Section 11.8) first introduced by Arthur Dempster in the 1960s and then brilliantly developed by Glenn Shafer in 1976. Since then this theory has been further developed by many, including Dempster (e.g., Dempster 2008) and a European school that includes work by Philippe Smets, Didier Dubois, and Thierry Denoeux (Denoeux et al. 2010).

11.7.1.1 *Representation through Basic Probability Assignment*

When $\mathcal{A}$ is a finite collection of events/sets, then a *belief function* can be defined in terms of what Shafer called a *basic probability assignment (bpa)*

$$m : \mathcal{A} \to [0,1],\ m(\emptyset) = 0,\ \sum_{A \in \mathcal{A}} m(A) = 1.$$

This function is like a probability mass function, with the important difference that it assigns weights to all *sets* in the algebra $\mathcal{A}$ (typically all subsets of Ω) rather than to just *elements* in Ω.

Definition 5 (Belief Function). *The belief function Bel corresponding to the bpa m is defined to be*

$$Bel(A) = \sum_{\{B : B \subseteq A\}} m(B).$$

For example, if $m(A) = 0$ whenever $\|A\| > 1$, then the resulting Bel is a standard probability measure.

11.7.1.2 *Basic Properties of Bel*

Lemma 2 (Properties of Bel). *It follows easily from the preceding definition and the properties of the bpa that*

$$(a)\ Bel(\emptyset) = 0,\ Bel(\Omega) = 1,\ A \supseteq B \implies Bel(A) \geq Bel(B),$$

$$(b)\ (\forall A, B \in \mathcal{A})\ Bel(A \cup B) + Bel(A \cap B) \geq Bel(A) + Bel(B).$$

Lemma (b) invokes the property of *2-monotonicity* that holds with equality for standard probability. Indeed, recall the *Inclusion-Exclusion Principle* expressing the standard probability μ of a union of any n events in terms of intersections of events,

$$\mu(\cup_{i=1}^{n} A_i) = \sum_{I \subseteq \{1,\ldots,n\}} (-1)^{1+I} \mu(\cap_{i \in I} A_i).$$

This result also holds for Bel with the important change of $=$ being replaced by $\geq$. In this sense, belief functions are the most probability-like of the extensions to standard probability.

11.7.1.3 *Vacuous Belief Function and Complete Ignorance*

The simplest example is the *vacuous* belief function defined by

$$\text{Bel}(A) = \begin{cases} 0 & \text{if } A \neq \Omega \\ 1 & \text{if } A = \Omega. \end{cases}$$

This vacuous belief function commits to nothing other than the sure event Ω, and, in this sense, it represents the epistemic state of *complete ignorance* more accurately than does, say, a uniform probability distribution when used to represent such ignorance on the classical probability account. The vacuous belief function is a far superior characterization of total ignorance than is a probability measure such as the uniform distribution.[27] The uniform probability distribution is a highly precise assertion about chance or uncertainty that does not differentiate between a state of exact knowledge of μ and one where μ is unknown. We will return to this issue of modeling ignorance in Section 11.9. Too often the highly precise model of a uniform probability distribution (say, in the case of a finite Ω) is taken to represent ignorance about the outcomes of a vaguely unspecified selection from Ω.

11.7.1.4 *Understanding the bpa m*

One can think of the bpa $m(A)$ as the amount of evidence or degree of belief focused exactly on A, or precisely supporting the occurrence or truth of event A *in toto*, and as not being subdivisible among the elements of A. In the words of Shafer (1976: p. 40)

> $m(A)$ measure the total portion of belief, or the total probability mass, that is confined to A yet none of which is confined to any proper subset of A.

The belief function Bel(A) can be understood as the minimum amount of evidence supporting A, and it is calculated as the sum of all of the basic probability assignments that precisely support the subsets of A. Alternatively, we can understand a belief function by expanding the two-valued logic of *true, false* to the ternary-valued *true, false, don't know*. Each set A corresponds to a triple $[p(A), q(A), r(A)]$ of non-negative terms that sum to one. $p(A)$ measures the weight of the evidence that directly supports the occurrence of A, $q(A)$ the weight of evidence directly supporting A^c, and $r(A)$ the extent to which we lack evidence.

[27] Or, for that matter, the refinement provided by the maximum entropy principle (see Jaynes 2003).

It follows that

$$r(A) = 1 - p(A) - q(A),\ q(A) = p(A^c),$$
$$\mathrm{Bel}(A) = p(A),\ \mathrm{Bel}(A^c) = q(A).$$

11.7.1.5 *Plausibility Function Pl*

We introduce the *plausibility function* $\mathrm{Pl}(A) = 1 - q(A)$ as all the possible evidence for A that is not contradicted by the evidence for A^c.

Definition 6 (Plausibility Function). *The plausibility function Pl is defined in terms of Bel by*

$$Pl(A) = 1 - Bel(A^c).$$

Lemma 3 (Plausibility Upper Bounds Belief).

$$(\forall A \in \mathcal{A})\ Pl(A) \geq Bel(A).$$

Proof. From the 2-monotonicity property of Bel and the definition of Pl

$$1 = \mathrm{Bel}(\Omega) \geq \mathrm{Bel}(A) + \mathrm{Bel}(A^c) \implies \mathrm{Pl}(A) = 1 - \mathrm{Bel}(A^c) \geq \mathrm{Bel}(A).$$

□

An axiomatic approach given in Section 11.8.3 will generalize these notions of belief and plausibility.

11.7.2 Belief Function Representation of CP Orders

Every CP ordering $\succsim$ can be represented by a belief function through

$$A \succsim B \iff \mathrm{Bel}(A) \geq \mathrm{Bel}(B),$$

although it is not the case that conversely every belief function represents a CP ordering. From the order-reversing property of complementation in CP noted in Section 11.6.2, if a specific Bel represents the CP $\succsim$ then so does its corresponding Pl,

$$A \succsim B \iff \mathrm{Pl}(A) \geq \mathrm{Pl}(B).$$

Bel is capable of representing more than just CP orders. When we say that an ordering of sets $\succsim$ *agrees with the partial ordering by set inclusion* we mean that

$$A \supseteq B \implies A \succsim B.$$

Similarly, a real-valued set function f agrees with the ordering of sets by inclusion when

$$A \supseteq B \iff f(A) \geq f(B).$$

The full representational scope of belief functions is characterized in Lemma 9 of Walley and Fine (1979) (also in Walley 1979), and is essentially (there are some qualifications) that the ordering agrees with the partial ordering by set inclusion.

As an illustration of a belief function representation of non-additive CP, we consider the non-additive KPS5 example of Kraft et al. (1959) for $\Omega = \{a, b, c, d, e, \}$. The four key inequalities leading to non-additivity (see Section 11.6.5) are

$$ac \prec d,\ ad \prec bc,\ cd \prec ae,\ be \prec acd.$$

Walley and Fine (1979: p. 364) provide an example of a belief function representation for KPS5

$$m(\{a\}) = 10/182,\ m(\{b\}) = 26/182,\ m(\{c\}) = 27/182,$$
$$m(\{d\}) = 40/182,\ m(\{e\}) = 60/182,$$
$$m(\{d, e\}) = 9/182,\ m(\{a, c, d\}) = 10/182.$$

It is readily verified that the multiplicative factor of $1/182$ is what is needed to have a sum of one for the basic probability assignments. Using this bpa for the KPS5 CP inequalities given above yields

$$182\text{Bel}(ac) = 37 < 182\text{Bel}(d) = 40,\ 182\text{Bel}(ad) = 50 < 182\text{Bel}(bc) = 53,$$
$$182\text{Bel}(cd) = 67 < 182\text{Bel}(ae) = 70,\ 182\text{Bel}(be) = 86 < 182\text{Bel}(acd) = 87,$$

in agreement with KPS5.

11.7.3 Mathematical Resolution in CP

On the one hand, where there exist agreeing probability measures and few equivalences between events, there is much latitude in choice of agreeing probability measures. For example, the particular 3-atom example of Section 11.6.4 has low mathematical resolution in that the CP order is compatible with a wide range of agreeing probability values (e.g., $1 > P(c) > 1/2$). On the other hand, with sufficiently many equivalences in a CP order there can be a unique agreeing probability measure. However, in this case, such a measure is restricted to having only rational values drawn from a limited set.

A different approach to resolution is needed when we have a non-additive CP order, such as the 5- or 6-element examples of Kraft et al. (1959), in which there are no equivalences between likelihoods of events and no agreeing probability measures. There is only a finite number, call it κ_n, of possible kinds of CP orders that are not equivalent under the $n!$ permutations (relabeling) of the n elements of Ω. The greater the choice measured by larger κ_n, the greater the potential mathematical resolution. There has been some research on estimating κ_n, the first such paper being that of Fine and Gill (1976) and a recent one by Conder et al. (2007). In any event, this very finiteness of the number of choices for CP orderings of the likelihoods of events implies a limited degree of resolution, particularly when compared to the infinite number of (even computable) probability measures on a sample space of even 2 elements.

11.8 Upper and Lower Unconditional Probability (U/LP)

11.8.1 Introduction to U/LP

We start with axioms common to a variety of U/LP formulations and refer specializations and their applicability to Cozman's *Handbook* article on "Imprecise and Indeterminate Probabilities" (Denoeux et al. 2010) and a remarkably original and thorough discussion in Walley (1991) that also treats upper and lower expectation. Let P^* denote *upper probability (UP)* and P_* denote *lower probability (LP)*. By introducing two real-valued functions, we may seem to be ignoring our opening arguments against the excessive precision of even a single real-valued standard probability. However, firstly, Axiom (iii) in Section 11.8.3 will show that either of the two functions can be defined in terms of the other one. Hence, there is only one independent function. Secondly, this pair of functions allows us now to be explicit (perhaps excessively so) about the precision or resolution of our U/LP representation through the gap $P^*(A) - P_*(A)$ for each event A. Thirdly, we are free to restrict ourselves to a finite set of possible numerical values for U/LP so as to better align the resulting limited precision of our U/LP model with the limited accuracy of the random source being modeled. This third option has the further advantage of avoiding the "paradox of the heap" that gives force to the example of intelligent life in the Andromeda galaxy that was described in Section 11.4.1.

In a subjectivist context we might think of $P^*(A)$ as the least upper bound for an individual's probability or credence that A occurs and P_* as a corresponding greatest lower bound; these bounds can be operational in terms of the individual's being unwilling to refine them further.[28]

In an objectivist context of repeated, unlinked random experiments we calculate the *relative frequencies* $\{r_n(A)\}$ that are the fraction of the first n experiments in which event A is observed to occur. We now allow for the possibility that the sequence of relative frequencies might not converge in the limit of n going to ∞.

To account for these possibly unending fluctuations, we introduce[29]

$$\limsup_{n\to\infty} r_n(A) \text{ and } \liminf_{n\to\infty} r_n(A).$$

For example, if

$$r_n(A) = a_n + b_n(-1)^n \text{ with } \lim_{n\to\infty} a_n = a > \lim_{n\to\infty} b_n = b > 0,\ a + b < 1,$$

$$\text{then } \limsup_{n\to\infty} r_n(A) = a + b < 1,\ \liminf_{n\to\infty} r_n(A) = a - b > 0.$$

[28] Walley goes further by discussing upper (E^*) and lower (E_*) expectations interpreted in terms of selling (least price we will accept) and buying (greatest price we will pay) prices for gambles. Joyce (2010) provides a defense for, and critical analysis of, imprecise credences and their use.

[29] *infimum (inf)* is a minimum when it exists and *supremum (sup)* is a maximum when it exists. Define $\liminf_{n\to\infty} r_n(A) = \lim_{n\to\infty} \inf_{m\geq n} r_n(A)$ and the same definition for limsup with "sup" replacing "inf".

Note that $\lim_{n\to\infty} r_n$ does not exist for this example because for any $\epsilon > 0$ there are infinitely many values of n for which the magnitude of $r_n - (a - b)$ is less than ϵ and infinitely many values of n for which the magnitude of $r_n - (a + b)$ is less than ϵ.

A limiting frequentist view, adapted to the possibility of divergent relative frequencies, is

$$P^*(A) = \limsup_n r_n(A) \text{ and } P_*(A) = \liminf r_n(A).$$

As r_n is always bounded between 0 and 1, these limits are well-defined even when relative frequencies do not converge as is required by the standard frequentist meaning of probability.

11.8.2 Agnosticism about Convergence of Relative Frequencies

In Section 11.9 we will use lower probability to establish a new possibility for an *agnostic* or *noncommital* stance about the long-run convergence of relative frequencies $\{r_n\}$, and thereby contribute to a very long-standing discussion (e.g., see Li and Vitanyi 2008, Sections 1.9, 2.4, 2.5, Jeffrey 1977, Hájek 1996) about justifying the meaning of standard probability as a convergent limit of relative frequencies. This result, not achievable within the confines of standard probability, demonstrates the greater useful expressivity of lower probability when compared with standard probability.

A common probability model for unlinked repetitions $\mathcal{E}_1, \ldots, \mathcal{E}_n, \ldots$ of a random experiment $\mathcal{E}$ is that they are mutually independent and identically distributed (i.i.d.). For relative frequencies $\{r_n(A)\}$ derived from the fraction of occurrences of an event A in the first n trials, it can be proven that this sequence converges in all of the usual senses of probabilistic convergence such as convergence in probability, in mean square, and with probability one. Importantly, the same results on convergence can also be shown to hold for a much larger collection of probability models known as *stationary random processes* (see Section 11.9.6) that also include the i.i.d. random processes.

Theorem 5 in Section 11.9.8 will assert that by widening the class of stationary probability models to include the class of stationary lower probability models, we can find instances of stationary lower probability models for which all we can say about the event C^* of convergence of relative frequencies is that $P_*(C^*) = 0$ and $P^*(C^*) = 1$—these models are completely agnostic about convergence. However, unlike the case of a vacuous belief function (see Section 11.7.1.3), these upper and lower probabilities can now take nondegenerate values for such archetypical observable events as cylinder sets (see Section 11.10).

11.8.3 Axioms for U/LP

Given an algebra of events $\mathcal{A}$ (possibly all subsets of the typically finite sample space Ω), then lower probability P_* and upper probability P^* are real-valued normalized set functions defined by the following axioms:

(i) (non-negativity) $(\forall A \in \mathcal{A}) P_*(A) \geq 0$;

(ii) (normalization) $P_*(\Omega) = 1$;

(iii) (conjugacy) $P^*(A) = 1 - P_*(A^c)$;

(iv) (superadditivity) $A \cap B = \emptyset$ implies $P_*(A \cup B) \geq P_*(A) + P_*(B)$;

(v) (subadditivity) $A \cap B = \emptyset$ implies $P^*(A \cup B) \leq P^*(A) + P^*(B)$.

Belief and plausibility functions satisfy these axioms, as does standard probability when the inequalities in Axioms (iv, v) are replaced by equalities. Axioms (i, ii) are counterparts of the same axioms for standard probability P. Axiom (iii) is new in that it introduces an upper probability companion P^* to a lower probability P_*—albeit, if you set the upper probability equal to the lower probability then you get a basic property of standard probability under set complementation. It is Axiom (iii) that shows we have introduced only one new real-valued function and not two; each is defined in terms of the other. Axioms (iv, v) generalize the key additivity property of standard probability to super- and sub-additivity for P_* and P^*, respectively. These axioms make sense when P_* and P^* are understood in the frequentist terms suggested above as well as in subjective terms that allow for either imprecision or indeterminacy in probability representations of degree of belief.

Lemma 4 (Basic Implications of the U/LP Axioms).

(a) $P_*(\emptyset) = 0$.

(b) $P^*(\emptyset) = 0$, $P^*(\Omega) = 1$.

(c) $A \supseteq B \implies P_*(A) \geq P_*(B)$ *and* $P^*(A) \geq P^*(B)$.

(d) Axiom (v) for P^* *is equivalent to* $P_*(A) + P_*(B) \leq 1 + P_*(A \cap B)$.

An important distinction between lower probabilities and probability measures is that a lower probability R_* defined on an algebra $\mathcal{A}$ of subsets of Ω, and satisfying the axioms on $\mathcal{A}$, can be extended to a lower probability P_* defined on all subsets of Ω through

$$(\forall B \subset \Omega)\ P_*(B) = \sup\{R_*(A) : A \subseteq B, A \in \mathcal{A}\}.$$

11.8.4 Domination, Envelopes, and 2-Monotone LP

Definition 7 (Dominated). *A* P_* *is dominated by a standard probability measure* μ *if for every event* $A \in \mathcal{A}$, $\mu(A) \geq P_*(A)$. *Otherwise,* P_* *is said to be undominated.*

A standard probability measure is trivially dominated by itself. For reasons given below, every belief function is dominated. If P_* is dominated by μ then P^* dominates μ in the obvious sense that $A \in \mathcal{A}$, $\mu(A) \leq P^*(A)$. In Section 11.9 we will make use of the fact that there exist undominated P_*.

Definition 8 (Envelopes). *Assume an arbitrary nonempty set* $\mathcal{M} = \{\mu\}$ *of standard probability measures on a common event algebra* $\mathcal{A}$. *For each* $A \in \mathcal{A}$, *define a lower envelope through* $P_*(A) = \inf_{\mu \in \mathcal{M}} \mu(A)$ *and an upper envelope through* $P^*(A) = \sup_{\mu \in \mathcal{M}} \mu(A)$.

It is immediate that every lower envelope is dominated by all of the measures in $\mathcal{M}$. Walley (1991) extensively discusses envelopes. Levi, Kyburg, Joyce, and others have used envelopes to represent credences and indeterminacy.

Relationships involving envelopes and 2-monotonicity are provided next.

1. It is not difficult to verify that the definition of envelopes satisfies Axioms (i)–(v).
2. Therefore, envelopes are also lower and upper probabilities.
3. When $\mathcal{M}$ contains only a single measure, $\|\mathcal{M}\| = 1$, the U/LP reduces to standard probability.
4. Less obviously, every belief function is also a lower envelope as a consequence of its being 2-monotone (see Section 11.7.1.2).
5. The limiting frequentist interpretation of U/LP given in Section 11.8.1, using $\liminf r_n(\cdot)$, $\limsup r_n(\cdot)$, is an option that yields lower and upper envelopes of probability measures corresponding to convergent subsequences of the relative frequencies.
6. It is the case that every 2-monotone lower probability and, therefore, every belief function is also a lower envelope.
7. However, it is false that a lower envelope need be 2-monotone (see Wolfenson 1979, Lemma 3.1 and p.50).
8. The use of envelopes in decision making is discussed in Wolfenson and Fine (1982) and Joyce (2010).

11.8.5 Mathematical Resolution for U/LP

The level of mathematical resolution of uncertainty, chance, or indeterminacy for an event A can be measured by the gap $\delta(A) = P^*(A) - P_*(A) \geq 0$. The worst-case level of mathematical resolution is measured by $\Delta = \max_A \delta(A)$. The vacuous belief/plausibility functions achieve the largest possible $\Delta = 1$. The larger is $\Delta \leq 1$, the poorer is the corresponding U/LP resolution. The source of the low resolution might be rooted either in the intrinsic indeterminacy of subjective beliefs, or in persistent fluctuations in long-run relative frequencies, or in contradictions in an epistemic propositional base.

11.9 Using Undominated LP to Be Noncommital About Tail Events Such as the Convergence of Relative Frequencies

11.9.1 Staking Out a Controversial Position

> Whereof one cannot speak, thereof one must be silent.
>
> (Wittgenstein 1922)

The celebrated *laws of large numbers*[30] of standard probability theory and their extensions to stationary and ergodic random processes[31] are brittle results whose utility deserves to

[30] See Theorem 4 in Section 11.9.3.
[31] See Section 11.9.7.

be questioned. These laws apply to situations where we contemplate an infinite sequence of random experiments (that may or may not be mutually independent) as is the case for, say, repeated unlinked performances of a given random experiment (e.g., the toss of a coin). Applications of the laws of large numbers enable us to evaluate the probabilities of so-called *tail events*[32] that extend over infinite time (e.g., limits of time averages of an infinite sequence of outcomes). However, whether or not a tail event occurs does not depend upon the occurrences of those individual component random experiments that can be observed in any given finite time. These tail events are intrinsically non-observable. The non-observability of tail events notwithstanding, the laws of large numbers make the strongest possible probabilistic assertions about the occurrence of such complex events that cannot be observed, even in principle.

Furthermore, while the strong laws of large numbers also give the tail events of probability either one or zero,[33] the smallest changes in assumed probabilities of the component random experiments can immediately lead to what were probability one tail events becoming probability zero tail events—thus the brittleness of these laws.

Tail events that are non-observable correspond to outcomes "whereof one cannot speak". The flexibility offered by upper and lower probability allows for a notion of vacuous events. Recall that an event A is *vacuous* with respect to P_* if $P_*(A) = 0, P^*(A) = 1 - P_*(A^c) = 1$. Asserting that A is vacuous is our equivalent of being "silent" about the occurrence of an event. The concept of being vacuous is absent from standard probability.[34] Precise numerical probability assignments violate the "silence" required by our ignorance of what will occur.

Theorem 5 of Section 11.9.8 will show that replacing such a standard probability model by appropriate U/LP models can enable us to closely match the standard probability model for cylinder set events that are determined by a specified finite number of outcomes of component random experiments[35] while also allowing us to remain noncommittal about the unknowable tail events. We thereby conform to Wittgenstein's dictum in that events whose outcomes cannot be determined in finite time correspond to matters "whereof one cannot speak" and access to the ability to be noncommittal corresponds to being able to "be silent".

11.9.2 Introduction to this Section

Hitherto, our discussion of standard probability and the advantages of its alternatives has focused on random experiments that are relatively simple. As discussed in Section 11.2, a *random experiment* $\mathcal{E}$ in standard probability is a triple $\mathcal{E} = (\Omega, \mathcal{A}, \mu)$, with μ being a standard probability measure assigning real numbers in the interval $[0, 1]$ to each set/event

[32] See Section 11.9.5.

[33] This is the content of the Kolmogorov Zero-One Law for tail events defined on infinite sequences of independent random experiments.

[34] In the confines of standard probability, ignorance (a version of being vacuous) about a finite collection of events is often held to be modeled by a very precise uniform standard probability distribution over the finitely many outcomes. Alternatively, we can determine an equally precise but different probability distribution by use of the maximum entropy principle (see Jaynes 2003).

[35] See Appendix.

$A \in \mathcal{A}$ in a manner that is consistent with the probability axioms. In probability modeling practice we talk more often about random variables than about events.

Definition 9 (Random Variable). *A* random variable *X is a real-valued (or vector-valued) function on Ω,*

$$X : \Omega \to \mathbb{R},$$

that satisfies a condition of $\mathcal{A}$-measurability. *Measurability means that for any $x \in \mathbb{R}$ the set*

$$\{\omega : X(\omega) \leq x\} \in \mathcal{A}.$$

We need to extend this setup to the case of *random or stochastic processes* that involve infinitely many random variables. The existence of countably infinitely many random variables is needed for us to define the set of tail events, mentioned in the previous section, and to carry out the argument that results in Theorem 5 of Section 11.9.8.

In order to provide concrete probabilistic models in support of our argument, we begin with the best known model of independent and identically distributed (i.i.d.) random variables. We then generalize to the much larger family of stationary discrete time random processes that have finite expected values. These latter standard probability random processes are then compared with U/LP random processes in Section 11.9.8; U/LP models can mimic the behavior of standard probability for observable events while also being vacuous or noncommital about events of which we cannot know whether or not they occur.

11.9.3 Processes of i.i.d. Random Variables

Jacob Bernoulli, working at the close of the 17th century, studied a basic random process generated by infinitely many repeated unlinked random experiments that gives rise to an infinite sequence of random variables $X_1, \ldots, X_n, \ldots$ that are identically distributed (same probability model). For example, we consider successive throws of a sturdy die whose physical composition is unaffected by being tossed repeatedly and whose nth outcome X_n takes values in $\{1, \ldots, 6\}$. We model unlinkedness by the formal concept of (stochastic) independence, with two events A, B (e.g., $X_1 = 3, X_5 = 4$) being independent if and only if their probabilities satisfy $P(A \cap B) = P(A)P(B)$. Let EX denote the common expected values[36] of the identically distributed random variables $\{X_i\}$. Bernoulli proved the first (weak) law of large numbers for the limiting behavior of the time averages of these random variables. Defining the *sample average*

$$S_n = \frac{1}{n}\sum_{1}^{n} X_i \text{ yields } (\forall \epsilon > 0) \lim_{n \to \infty} P(|S_n - EX| > \epsilon) = 0.$$

This result was subsequently strengthened to the

[36] If X is a discrete random variable taking the value x_k with probability p_k, then $EX = \sum_k x_k p_k$ when this sum is defined.

Theorem 4 (Kolmogorov Strong Law of Large Numbers). *If the random variables* $\{X_i\}$ *are i.i.d. with common expected value* EX *and* $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, *then*

$$P\left(\lim_{n\to\infty} S_n = EX_1\right) = 1.$$

Restated, the sequence of time averages converges to EX *with probability one.*[37]

While Theorem 4 seems to settle the issue of the long-run behavior of i.i.d random variables, observe that for every finite n, the particular sample average S_n has no influence whatsoever on the limiting behavior of the infinite sequence of sample averages $\{S_m, m \in \mathbb{N}\}$. This is in contrast to, say, the evaluation of an infinite sum of random variables where the first n summands do make a contribution to the total sum when it is finite. The theorems are theorems, but they may deserve less prominence than they have been given. Recall John Venn's caution (Section 11.4.2) that, in reality, random experiments if repeated long enough will cease to exhibit such behavior because of unavoidable and unpredictable changes in the experimental apparatus and its environment. Thus, according to Venn, the above two celebrated theorems are inapplicable in practice when the sample size n becomes large enough. Furthermore, from the viewpoint of Section 11.9.2, the independence of the first $X_1, \ldots, X_n$ from $X_{n+1}, X_{n+2}, \ldots$ means that S_n is asymptotically independent of S_m with increasing $m >> n$. Arithmetic tells us that we cannot learn $\lim_{m\to\infty} S_m$ from any finite initial sequence of values of the relative frequencies, however long the finite sequence may be.

Our position is that you should not take seriously any very long-range prediction. However, you may place confidence in the stability or reliability of short- or medium-term predictions. In the subsections to follow we generalize from the i.i.d. case to the stationary case and then provide Theorem 5 in Section 11.9.8 showing that the introduction of U/LP achieves nearly the same results as any standard probability model for events that are determined by a given number of random variables. However, for the tail events (see Section 11.9.5) involving infinitely many random variables and defined through limiting properties, our U/LP model will be vacuous to reflect your being noncommittal, ignorant, or agnostic about such non-observable events.

Section 11.9.8 provides a large family of examples that includes the asymptotic behavior of the time averages of outcomes of infinitely repeated random experiments. Sections 11.9.4, 11.9.5, and 11.9.7, together with the Appendix, provide compact reviews of the relevant elements of the standard probability theory of discrete-time random processes. This background is needed for you to understand the contribution in Section 11.9.8 made possible by replacing standard probability by U/LP.

11.9.4 Defining a Discrete-Time Random Process

A random variable X (see Definition 9 or Fine 2006, Section 8.2), defined for a random experiment $\mathcal{E} = (\Omega, \mathcal{A}, \mu)$, is described by a *cumulative distribution function (cdf)*

$$F_X(x) = \mu(\{\omega : X(\omega) \leq x\}).$$

[37] To avoid being overwhelmed by the seeming strength of this conclusion, it helps to recall that while there is an uncountable number of possible infinite sequences for which convergence holds, there is also an uncountable number of infinite sequences for which it does not hold.

This definitional process extends, as well, to several random variables through their joint cdf.

We focus on a *discrete time random process*[38] $\mathbb{X}$ that may model such temporal sequences indexed by time t, as the amount of rain falling during hour t in a specified geographic region; the maximum temperature reached in that region on day t; the outcome of the t-th spin of a (very durable!) roulette wheel; the daily maximum of a stock price index; or the waiting time for service of the t-th internet packet arriving at an internet node. In the case of rainfall and the stock index we would expect statistical dependence between successive values over a time interval $|t_i - t_j|$ that is not large.

Definition 10 (Random Process). *A discrete time random process $\mathbb{X}$ is an indexed by $\mathbb{Z} = \{0, \pm 1, \pm 2, \ldots\}$ collection of random variables $\{X_t, t \in \mathbb{Z}\}$, each taking values in $\mathcal{X} \subseteq \mathbb{R}$. The collection of random variables is described by a probability measure $\mu_{\mathbb{X}}$ defined on the σ-algebra*[39] $\mathcal{A} = \sigma(\cdots, X_{-1}, X_0, X_1, \cdots)$ *of events generated by these random variables.*

In this framework, given $\mathcal{X}$, any integer $k \geq 1$ and any $t_1, \ldots, t_k \in \mathbb{Z}$ we can consistently specify the multivariate cdf

$$(\forall x_1, \ldots, x_k \in \mathbb{R})\ F_{X_{t_1}, \ldots, X_{t_k}}(x_1, \ldots, x_k) = \mu_{\mathbb{X}}(X_{t_1} \leq x_1, \ldots, X_{t_k} \leq x_k),$$

in terms of $\mu_{\mathbb{X}}$ and, conversely, specify $\mu_{\mathbb{X}}$ from an infinite consistent collection of such cumulative distribution functions.

11.9.5 Tail Events

Given any random process $\mathbb{X}$ of random variables indexed by $\mathbb{Z}$ we identify the collection $\mathcal{T}$ of *tail events*, events for each of which knowledge of the values of any finite collection of random variables provides no information whatsoever about the outcomes of any tail event in $\mathcal{T}$. $\mathcal{T}$ is called the *tail σ-algebra* of events. We focus on these events in Section 11.9.8.

Definition 11 (Tail σ-Algebra of Events). *For any n consider the σ-algebra $\mathcal{T}_n = \sigma(X_n, X_{n+1} \ldots)$, of events determined only by random variables at times at least as large as n. Define the* tail σ-algebra

$$\mathcal{T} = \cap_{n=1}^{\infty} \mathcal{T}_n = \cap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, \ldots).$$

Thus $\mathcal{T}$ is a σ-algebra containing events whose outcomes do not depend upon any $X_t, t \leq n$ for any finite n. No finite number of random variables provides any information about an event in $\mathcal{T}$. An example of such an event C^* in $\mathcal{T}$ is that of convergence of sample averages,

$$C^* = \{\omega : \lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) = X(\omega)\}.$$

[38] There are numerous texts and monographs on this subject that are written at a variety of levels of mathematical sophistication. For definiteness I cite Fine (2006), chapter 20.

[39] See Footnote 2.

Knowledge of the values of any $X_{t_1} = x_i, \ldots, X_{t_k} = x_k$ provides no information about the random quantity C^*. The complementary event

$$D^* = \{\omega : \limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega) > \liminf_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} X_j(\omega)\},$$

of divergence of sample averages, is also in $\mathcal{T}$.

> Events in $\mathcal{T}$ are not observable as their occurrence is not influenced by any finite sequence of random variables. The members of $\mathcal{T}$ fit a description of "whereof one cannot speak".

Nonetheless, when the random variables are independent (not necessarily identically distributed) the *Kolmogorov Zero-One Law* (see Billingsley 1979) assures us that tail events have probability either one or zero—we are "almost sure" about whether or not these unknowable events occur.

11.9.6 Defining a Stationary Random Process

We generalize from i.i.d. random processes to the much larger collection of *stationary random processes* for which Theorem 5 of Section 11.9.8 will hold. Stationary random processes are those where the underlying, say, physical mechanisms generating the random process do not vary with a choice of time origin—there is no temporal evolution in their model. For example, for any t_i, t_j, X_{t_i} and X_{t_j} have the same probability description

$$(\forall t_i, t_j \in \mathbb{Z})(\forall x \in \mathbb{R})\ F_{X_{t_i}}(x) = F_{X_{t_j}}(x).$$

More generally, the joint cdf

$$F_{X_{t_1},\ldots,X_{t_k}}(x_1, \ldots, x_k) = \mu_{\mathbb{X}}(X_{t_1} \leq x_1, \ldots, X_{t_k} \leq x_k)$$

is such that for any $s \in \mathbb{Z}$ we also have

$$F_{X_{t_1},\ldots,X_{t_k}}(x_1, \ldots, x_k) = F_{X_{s+t_1},\ldots,X_{s+t_k}}(x_1, \ldots, x_k).$$

However, since we need a definition of a stationary random process that will apply as well to U/LP, we use an approach equivalent to that just described but one not using cdfs. We still have a set $\mathcal{X}$ in which the random variables $X_i, i \in \mathbb{Z}$ each take their values. Let $x_i \in \mathcal{X}$ denote a value of X_i. An infinite sequence of values for all $i \in \mathbb{Z}$ is then denoted by

$$\underline{x} = \{\cdots, x_{-1}, x_0, x_1, \cdots\} \in \mathcal{X}^{\infty},$$

and is a possible history of all of the outcomes of the infinitely many random variables. Introduce the *left shift operator*

$$T : \mathcal{X}^{\infty} \to \mathcal{X}^{\infty},\ (\forall i \in \mathbb{Z})(\forall \underline{x} \in \mathcal{X}^{\infty})\ T\underline{x} = \{(T\underline{x})_i = x_{i+1}\},\ TA = \{T\underline{x} : \underline{x} \in A\}.$$

Correspondingly, there is also a *right shift operator*

$$T^{-1} : \mathcal{X}^{\infty} \to \mathcal{X}^{\infty},\ (\forall i \in \mathbb{Z})(\forall \underline{x} \in \mathcal{X}^{\infty})\ (T^{-1}\underline{x})_i = x_{i-1}.$$

Definition 12 (Closure Under a Time Shift). *The σ-algebra $\mathcal{A}$ is* closed *under T, T^{-1} if*

$$(\forall A \in \mathcal{A})\ TA \in \mathcal{A}, T^{-1}A \in \mathcal{A}.$$

Definition 13 (Invariant Event and Set). *A set $A \in \mathcal{I}$ is invariant under T if $A = TA = T^{-1}A$. If this holds for all sets in $\mathcal{A}$, then $\mathcal{I}$ is a T-invariant set.*

The tail σ-algebra $\mathcal{T}$ is an invariant set under all T, T^{-1}.

Definition 14 (Stationary Set Function). *If a set function ρ has domain $\mathcal{A}$ that is closed under all T, T^{-1}, then ρ is a stationary set function if*

$$(\forall T)(\forall A \in \mathcal{A})\ \rho(A) = \rho(TA) = \rho(T^{-1}A).$$

Definition 15 (Stationary Random Process). *A random process $\mathcal{X}, \mathbb{Z}, \mathbb{X}, \mathcal{A}, \mu_{\mathbb{X}}$ is stationary if $\mathcal{A}$ is closed under all time shifts T, T^{-1} and*

$$(\forall A \in \mathcal{A})\ \mu_{\mathbb{X}}(A) = \mu_{\mathbb{X}}(TA) = \mu(T^{-1}A).$$

Stationary random processes are widely discussed (e.g., see Fine 2006, Section 20.5, Fristedt and Gray 1997, chapter 28).

If a random process is non-stationary (its statistics varying with the choice of time origin), then it will not be surprising if it fails to have convergent relative frequencies. Hence, we focus on (the commonly used idealization of) stationary random processes.

We can now define the extension of stationarity to lower and upper probability through

Definition 16 (Stationary Upper/Lower Probability). *A lower probability P_* and upper probability P^* are stationary (invariant under time shift T) if for every event A and shift T*

$$P_*(A) = P_*(TA),\ P^*(A) = P^*(TA).$$

We could also include invariance under T^{-1} but it will not be needed. For U/LP we will take $\mathcal{A}$ to be the power set (all subsets) of Ω. Hence, closure of $\mathcal{A}$ under T, T^{-1} is guaranteed.

11.9.7 Laws of Large Numbers for Stationary Standard Probability

In standard probability, a stationary random process $\mathbb{X}$ with finite expected value EX_o (equals any EX_k) obeys the celebrated *stationarity convergence* and *ergodic* theorems (e.g., see Fristedt and Gray 1997: p. 558 or Loève 1977: p. 87) that ensure that, say, long-run (i.e., asymptotically as the time index in $\mathbb{Z}$ goes to infinity) time averages of outcomes $\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\}$ will converge with probability one to a random variable (stationarity convergence),

$$\mu_{\mathbb{X}}(C^*) = 1,\ \mu_{\mathbb{X}}(D^*) = 0.$$

> When we focus our attention on a stationary discrete-time random process, with random variables having finite expected value, then standard probability does not allow us to suspend judgment about the existence, either with probability one or with probability zero, of limits of sample averages. Standard probability "speaks" when it should remain "silent".

While the tail event of convergence to a limit is not observable, nevertheless we are not silent about its existence as a random variable limit in the general stationary case or silent about its value being EX_o in the ergodic case.

11.9.8 The Main Theorem

We assert that upper/lower probability models can be noncommittal, vacuous, or agnostic about tail events whereas standard probability is forced (by the theorems cited above) to make a precise commitment. Yet, upper/lower probability models can still closely approximate these standard probability models when we focus on cylinder sets[40] of diameter[41] no larger than a chosen value. By "noncommittal" or "agnostic" about an event A we mean that A is vacuous with respect to a given P_*. Unlike the case of standard probability, where every event receives a precise numerical probability, upper/lower probability allows us to maximally suspend judgment—to remain silent—about the occurrence of A. A collection of events is vacuous if each event in the collection is vacuous.

The following theorem is stated and proven in Sadrolhefazi and Fine (1994) as Proposition 5.8.

Theorem 5 (Existence of Continuous Upper/Lower Probabilities That Are Noncommittal About Tail Events). *Given any stationary lower probability*[42] R_**, with corresponding upper probability* R^**, an integer* $N \geq 1$ *and* $0 < \epsilon < 1$*, there exists a stationary*[43] *and monotonely continuous*[44] *lower probability* P_**, with corresponding* P^**, that is vacuous on the set* $\mathcal{T}$ *of tail events*[45] *and yet is similar to* R_* *in that for all cylinder sets* C *of diameter*[46] *no greater than the chosen* N*,*

$$|R_*(C) - P_*(C) \leq \epsilon \text{ and } |R^*(C) - P^*(C) \leq \epsilon.$$

Theorem 5 asserts the startling fact that, from a non-asymptotic viewpoint, you are able to arbitrarily closely approximate to any given stationary lower probability R_* (e.g., even a standard i.i.d. probability measure) by a stationary P_* that is agnostic about the long-run convergence of relative frequencies and is regular in the sense that it is also monotonely continuous along the full collection $\mathcal{C}$ of cylinder sets. This approximation can be within any desired distance $\epsilon > 0$ between R_* and P_*. The role of a given $N \geq 1$, is that this

[40] See Appendix.
[41] See Appendix.
[42] See Definition 16. We can allow R_* to be a standard probability measure.
[43] See Definition 15.
[44] Monotone continuity along the cylinder sets in $\mathcal{C}$, as defined for $\mathcal{A}$ in Section 11.2.
[45] See Definition 11.
[46] See Definition 18.

approximation to within ϵ is guaranteed only for cylinder sets having a diameter (see Definition 18) no greater than the chosen N. Restated, you arbitrarily closely approximate any stationary R_* by, say, P_* on those cylinder set events whose outcomes are determined by finitely many (no more than the diameter) random variables and their occurrence or non-occurrence is determined by waiting no longer than the diameter from the onset of the occurrence of the events. Achieving this result may require that P_* be a lower probability that is undominated[47] and therefore has little relationship to any standard probability measure.

> There exist monotonely continuous, stationary lower probabilities in which the event C of convergence of relative frequencies has $P_*(C) = 0$ and $P^*(C) = 1$. This provides a precise meaning for your agnosticism: your confidence in the event of convergence has a lower probability of zero and a corresponding upper probability of one, yielding a resolution width $\Delta = 1$ that is as large (imprecise) as it can be.

> Lower probability allows you to remain noncommittal ("silent", vacuous, agnostic) about the long-run convergence of sample averages (e.g., relative frequencies) while also mimicking (to within preassigned ϵ) the behavior of any other stationary lower probability on all those observable events whose realizations are determined within a prespecified limited time span.

11.10 Envoi

We motivated the need for alternative mathematical probability models in Sections 11.1, 11.3, and 11.4. The alternative mathematical concepts of "possible", "probable", and "at least as probable as" were defined and explored in Sections 11.5 and 11.6. In each case we noted the partial connections of these approaches to standard probability and found that the latter failed to address all instances of the former. We introduced belief functions in Section 11.7 as an extension from standard probability and as a general representation for comparative probability orderings, and then commented on why belief functions are of interest in their own right. In Section 11.8, we extended from belief functions to the larger class of upper and lower probabilities. As many aspects of upper and lower probabilities are addressed by Fabio Cozman in Chapter 14 of this volume, in Section 11.9 we focused on their novel ability to allow us to remain noncommittal about such tail events as the long-run convergence of sample averages and relative frequencies. We should not be forced to assign probabilities to events either about which nothing is known or whose occurrences are in principle unobservable.

Walley (1991) provides a thoroughly reasoned examination of upper and lower probability, their extensions to upper and lower expectations, and their uses in inference. Halpern (2003) uses methods of mathematical logic to define and examine several generalizations of real-valued probability that are suited to reasoning about uncertainty.

Knowing about and developing alternative mathematical models of chance and uncertainty, that have various levels of resolution or precision, can give you access to mathematical models that more faithfully match the true range of levels of accuracy you encounter in empirical domain (real-world) phenomena of chance, uncertainty, and indeterminacy.

[47] See Definition 7 of Section 11.8.4.

Acknowledgments

I gratefully acknowledge the helpful comments by the *Handbook* co-editors, Alan Hájek and Chris Hitchcock, and the reviewer John Cusbert.

Appendix

Cylinder Sets

Definition 17 (Cylinder Sets). *A* cylinder set *C is an event whose occurrence is determined by a finite positive integer k, time points* $t_1, \ldots, t_k$*, random variables* $X_{t_1}, \ldots, X_{t_k} \in \mathbb{X}$*, a subset* C' *of the k-dimensional real numbers* $\mathbb{R}^k$*, and C occurs if and only if* $(X_{t_1}, \ldots, X_{t_k}) \in C'$*,*

$$C = \{\omega \in \Omega : (X_{t_1}(\omega), \ldots, X_{t_k}(\omega)) \in C'\} \in \mathcal{A}.$$

$\mathcal{C}$ *denotes the collection of all cylinder sets in* $\mathcal{A}$*.*

Hence, the occurrence or non-occurrence of the above-described *C* is observable by time t_k.

Definition 18 (Dimension and Diameter). *A cylinder set C, as described in Definition 17, has* dimension *k and* diameter $t_k - t_1 + 1$.

References

Balding, D. (2011) Book review: Taroni, F., Aitken, C., Garbolino, P. and Biedermann, A. (2006) Bayesian Networks and Probabilistic Inference in Forensic Science. *Law, Probability Risk*. 10. pp. 355–8.

Billingsley, P. (1979) *Probability and Measure*. p. 247. New York, NY: Wiley.

Black, M. (1967) Probability. In Edwards, P. (ed.). *The Encyclopedia of Philosophy*. 6. New York: Macmillan Publishing Co. Reprinted in 1970.

Black, M. (1970) *Margins of Precision: Essays in Logic and Language*. Epigraph. Ithaca, NY: Cornell University Press.

Brun, W., Teigen, K. (1988) Verbal probabilities: ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes*. 41. pp. 390–404.

Budescu, D. and Wallsten, T. (1995) Processing linguistic probabilities: general principles and empirical evidence. *Psychology of Learning and Motivation*. 32. pp. 275–318.

Budescu, D., Broomell, S. and Por, H.-H. (2009) Improving communication of uncertainty in the reports of the Intergovernmental Panel on Climate Change. *Psychological Science*. 20.3. pp. 299–308.

Cahan, A., Gilon, D., Manor, O., and Paltiel, O. (2005) Clinical experience did not reduce the variance in physicians' estimates of pretest probability in a cross-sectional survey. *J. of Clinical Epidemiology*. 58. pp. 1211–16.

Caves, C., Fuchs, C. and Schack, R. (2002) Quantum probabilities as Bayesian probabilities. *Physical Review A*. 65. 022305.

Conder, M., Searles, D. and Slinko, A. (2007) Comparative probability orders and the flip relation. In de Cooman, G., Vejnarova, J., Zaffalon, M. (eds.). *ISIPTA '07 Proc. Fifth Inter. Symp. on Imprecise Probability: Theories and Applications*. SIPTA. pp. 67–76.

de Finetti, B. (1931) On the subjective meaning of probability. *Fundamenta Mathematica*. 17. pp. 298–329. (In Italian).

de Finetti, B. (1990) *Theory of Probability: A critical introductory treatment*. Volume 1. Translated by Macchi, A. and Smith, A. Chichester and New York, NY: Wiley.

Dempster, A. (2008) The Dempster-Shafer calculus for statisticians. *International Journal of Approximate Reasoning*. 48. pp. 365–77.

Denoeux, T., Younes, Z., and Abdallah, F. (2010) Representing uncertainty on set-valued variables using belief functions. *Artificial Intelligence*. 174. pp. 479–99.

Emmerton, J. (c. 2002) Birds' judgments of number and quantity. Avian Visual Cognition Program, Psychology Department, Tufts University. [Online] Available from: http://www.pigeon.psy.tufts.edu/avc/emmerton [Accessed 8 Aug 2015].

Eriksson, L. and Hájek, A. (2007) What are degrees of belief? *Studia Logica*. 86. pp. 183–213.

Fine, T. (1973) *Theories of Probability: An Examination of Foundations*. Walthan, MA: Academic Press.

Fine, T. (1977) An argument for comparative probability. In Butts, R. E. and Hintikka, J. (eds.) *Basic Problems in Methodology and Linguistics*. pp. 105–19. Dordrecht: D. Reidel.

Fine, T. L. (1994) Upper and lower probability. In Humphreys, P. (ed.) *Patrick Suppes: Scientific Philosopher, I*. Synthese Library 233. pp. 109–33. Dordrecht: Kluwer.

Fine, T. L. (2006) *Probability and Probabilistic Reasoning for Electrical Engineering*. Upper Saddle River, NJ: Pearson/Prentice-Hall.

Fine, T. and Gill, J. (1976) The enumeration of comparative probability relations. *The Annals of Statistics*. 4. pp. 667–73.

Finkelstein, M. and Fairley, W. (1971) The continuing debate over mathematics in the law of evidence: a comment on "Trial by Mathematics". *Harvard Law Review*. 84. 8. pp. 1801–9.

Fristedt, B. and Gray, L. (1997) *A Modern Approach to Probability Theory*. Boston, MA: Birkhauser.

Google (2012) The self-driving car logs more miles on new wheels. [Online] Available from: http://googleblog.blogspot.com/2012/08/the-self-driving-car-logs-more-miles-on.html [Accessed 8 Aug 2015].

Gorry, G. A., Kassirer, J., Essig, A., and Schwartz, W. (1973) Decision analysis as the basis for computer-aided management of acute renal failure. *The American J. of Medicine*. 55. pp. 473–84.

Greenhalgh, Trisha (2010) *How To Read a Paper: The Basics of Evidence-Based Medicine*. 4th ed. (chapter 1, section 1.1.) Oxford: Wiley-Blackwell.

Hájek, A. (1996) "Mises redux"–redux: fifteen arguments against finite frequentism. *Erkenntnis*. 45. 2/3. pp. 209–27.

Hájek, A. (2008) Arguments for–or against–probabilism. *Brit. J. Philosophy of Science*. 59. pp. 793–819.

Hájek, A. and Hitchcock, C. (eds.) (2016) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Halpern, J. Y. (2003) *Reasoning about Uncertainty*. Cambridge, MA: MIT Press.

Hamblin, C. (1959) The modal "probably". *Mind*. 68. pp. 234–40.

Iglehart, J. K. (2005) Putting evidence into practice. *Health Affairs*. 24. 1.

IPCC (2005) *IPCC Guidance Notes for Lead Authors of the IPCC Fourth Assessment Report on Addressing Uncertainties*. [Online] Available from: http://www.ipcc.ch/meetings/ar4-workshops-express-meetings/ uncertainty-guidance-note.pdf. [Accessed 8 Aug 2015].

IPCC (2014) Plans for Completion of Fifth Assessment Report (AR5) by October 2014. [Online] Available from: http://www.ipcc.ch/activities/activities.shtml#.T5BYlI6uHnA [Accessed 8 Aug 2015].

Jaynes, E. Posthumously edited by Bretthorst, G. L. (2003) *Probability Theory: The Logic of Science*. See p. 345 and sections 11.6, 12.1, 12.2. Cambridge and New York: Cambridge University Press.

Jeffrey, R. (1977) Mises redux. In Butts, R. E. and Hintikka, J. (eds.) *Basic Problems in Methodology and Linguistics*. pp. 213–22. Dordrecht, Boston: D. Reidel.

Joyce, J. (2010) A defense of imprecise credences in inference and decision making. *Philosophical Perspectives*. 24. Epistemology. pp. 281–323.

Kaplan, M. A. and Fine, T. L. (1977) Joint orders in comparative probability. *The Annals of Probability*. 5. 161–79.

Keynes, J. M. (1921) *A Treatise on Probability*. London: Macmillan.

Kolmogorov, A. N. (1933/1956) *Foundations of the Theory of Probability*. Translated from the German by Nathan Morrison. New York, NY: Chelsea Publishing Company.

Kraft, C., Pratt, J., and Seidenberg, A. (1959) Intuitive probability on finite sets. *Annals of Math. Statistics*. 30. pp. 408–19.

Krantz, D., Luce, R. D., Suppes, P., and Tversky, A. (1971) *Foundations of Measurement* Vol. I*: Additive and polynomial representations*. New York, NY: Academic Press.

Kumar, A. (1982) *Lower Probabilities on Infinite Spaces and Instability of Stationary Sequences*. Ph.D. dissertation. Ithaca, NY: Cornell University.

Kyburg, H. (1974) *The Logical Foundations of Statistical Inference*. Dordrecht-Holland and Boston, MA: D. Reidel.

Levi, I. (1985) Imprecision and indeterminacy in probability judgment. *Philosophy of Science*. 52. pp. 390–409.

Li, M. and Vitanyi, P. (2008) *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd ed. New York, NY: Springer.

Loève, M. (1977) *Probability Theory II*. 4th ed. New York, NY: Springer-Verlag.

Luce, R. D. (1996) The ongoing dialog between empirical science and measurement theory. *J. of Mathematical Psychology*. 40. pp. 78–98.

Maher, P. (2006) Book review: David Christensen, *Putting Logic in its Place: Formal Constraints on Rational Belief. Notre Dame Journal of Formal Logic*. 47. pp. 133–49.

Mode, E. (1963) Probability and criminalistics. *JASA*. 303. pp. 628–40.

Narens, L. (2007) *Theories of Probability: An examination of logical and qualitative foundations*. Singapore and Hackensack, NJ: World Scientific.

Neumann, J. von (1947) The mathematician, Parts I, II. *Works of the Mind*. 1. pp. 180–96. Chicago, IL: Univ. of Chicago Press.

Pauker, S. and Kassirer, J. (1980) The Threshold approach to clinical decision making. *New England J. of Medicine*. 302. 20. pp. 1109–17.

Postkasse (2012) Here are your cars speedometer error margin [sic]. [Online] Available from: http://click2how.com/your-speedometer-wrong-can-drive-faster/ [Accessed 8 Aug 2015].

Reagan, R. T., Mosteller, F. and Youtz, C. (1989) Quantitative meanings of Verbal Probability Expressions. *J. of Applied Psychology*. 74. pp. 433–42.

Regoli, G. (2000) Comparative probability orderings. In SIPTA Documentation on Imprecise Probability. [Online] Available from: http://www.sipta.org/documentation, paper by Regoli. [Accessed 8 Aug 2015].

Reznikova, Z. and Ryabko, B. (2011) Numerical competence in animals, with an insight from ants. *Behaviour*. 148. pp. 405–34.

Sadrolhefazi, A. and Fine, T. L. (1994) Finite-dimensional distributions and tail behavior in stationary interval-valued probability models. *Annals of Statistics*. 22. pp. 1840–70.

Sagan, C. (1973) *Communication with Extraterrestrial Intelligence (CETI)*. Cambridge, MA: MIT Press.

Scott, D. (1964) Measurement structures and linear inequalities. *J. Mathematical Psychology*. 1. 233–47.

Shafer, G. (1976) *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton Univ. Press.

Shafer, G. and Vovk, V. (2006) The sources of Kolmogorov's *Grundbegriffe*. *Statistical Science*. 21. 1. pp. 70–98.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., and Goodman, N. D. (2011) How to grow a mind: statistics, structure, and abstraction. *Science*. 331. 11 March 2011. pp. 1279–85.

Tillers, P. (2011) Trial by mathematics—reconsidered. *Law, Probability and Risk*. 10. 167–73.

Tribe, L. (1971) The continuing debate over mathematics in the law of evidence: a further critique of mathematical proof. *Harvard Law Review*. 84. 8. pp. 1810–20.

Venn, John (1888) *The Logic of Chance*. 3rd ed. New York, NY: Chelsea Publishing Company.

Walley, P. (1979) *Varieties of Modal and Comparative Probability*. Ph.D. dissertation. Ithaca, NY: Cornell University.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London and New York, NY: Chapman and Hall. See especially chapter 2, "Coherent Previsions".

Walley, P. and Fine, T. (1979) Varieties of modal (classificatory) and comparative probability. *Synthese*. 41. pp. 321–74.

Wang, Ucilla (2014) Driverless cars are data guzzlers. *Wall Street Journal*. 23 Mar. 2014.

Wittgenstein, L. (1922) *Tractatus Logico-Philosophicus*. Translated fron the German by C. K. Ogden. Proposition 7. London: Routledge.

Wolfenson, M. (1979) *Inference and Decision Making Based on Interval-Valued Probability*. Ph.D. dissertation. Ithaca, NY: Cornell University.

Wolfenson, M. and Fine, T. (1982) Bayes-like decision making with upper and lower probabilities. *Jour. American Statistical Assoc.* 77. pp. 80–88.

Yalcin, S. (2010) Probability operators. *Philosophy Compass*. 5. pp. 916–37.

CHAPTER 12

PROBABILITY AND NONCLASSICAL LOGIC

J. ROBERT G. WILLIAMS

Classical tautologies have probability 1. Classical contradictions have probability 0. These familiar features reflect a connection between standard probability theory and classical logic. In contexts in which classical logic is questioned—to deal with the paradoxes of self-reference, or vague propositions, for the purposes of scientific theory or metaphysical anti-realism—we must equally question standard probability theory.[1]

Section 1 covers the intended interpretation of 'nonclassical logic' and 'probability'. Section 2 reviews the connection between classical logic and classical probability. Section 3 briefly reviews salient aspects of nonclassical logic, laying out a couple of simple examples to fix ideas. Section 4 explores modifications of probability theory. The variations laid down will be motivated initially by formal analogies to the classical setting. In section 5, however, two foundational *justifications* are reviewed for the presentations of 'nonclassical probabilities' that are arrived at. Sections 6-7 describe extensions of the nonclassical framework: to conditionalization and decision theory in particular. Section 8 will consider some alternative approaches, and section 9 evaluates progress.

12.1 Preliminaries

Our topic is the interaction between nonclassical logic and probability. But 'nonclassical logic' and 'probability' in what sense?

In the following sections, we operate with a fairly narrow understanding of nonclassicality. For present purposes a nonclassical logic is one that diverges from classical orthodoxy on which arguments (sequents) or inferential rules are valid or invalid. An example of a

[1] For the paradoxes of self-reference, Field (2008) provides a recent survey of nonclassical approaches. For vagueness, see inter alia Williamson (1994); Keefe (2000); Smith (2008). Hughes (1992) is a relatively accessible approach to the issues surrounding quantum logic, with chs. 7 and 8 particularly pertinent to our concerns. For metaphysical anti-realism and logic, a locus classicus is Dummett (1991).

classically valid argument might be disjunctive syllogism: $A, \neg A \vee B \models B$. A rule of inference (in the technical sense that contrasts with this) involves a transition from one validity to another. Conditional proof is an example: this tells us that if $A, B \models C$ then also $A \models B \rightarrow C$. Nonclassical logics might declare one or the other, or both, invalid. For our purposes, the sentences in question can (usually) be thought of as drawn from a standard propositional language containing negation, disjunction, conjunction, and a conditional.

A second sense of nonclassicality pertains to semantics. A theory is nonclassical in this sense if it diverges from classical orthodoxy on what truth statuses there are or how they can be distributed. Classical semantics endorses bivalence—every meaningful sentence is either true or false. A nonclassical semantics may allow for sentences which are neither true nor false; or for intermediate *degrees* of truth. This is not nonclassicality in logic, strictly speaking; but the two forms of nonclassicality are intimately related and I will discuss both.[2]

A broader reading of 'nonclassical logic' would include the logic of operators and connectives beyond the usual propositional list. Modal, temporal, and conditional logics are paradigms of this kind of departure from classicality.[3] This broad reading won't be the focus of our discussion, but the interested reader is directed to the result from Paris (2001) quoted in Section 3. To a first approximation, the result shows that unless we have nonclassicality in the narrow sense, the theory of probability is unchanged (a remarkable result in its own right!).

Like nonclassicality, what we call 'probability' can vary along several dimensions. The first dimension of variation is the *kind* of phenomenon in question—perhaps rational belief, or objective chance, or degree of evidential confirmation.[4] In the sections below we focus on a *subjective* interpretation of probability. On this picture, subjects have beliefs that come in degrees. The *probabilist* then maintains that to be ideally rational, the distribution of these degrees of belief must be *probabilistic*—i.e., satisfy the probability axioms. Alternative interpretations are considered in the penultimate section of this survey.

A second dimension of variation concerns the items that probabilities attach to. We can choose between investigating probabilities that attach to fine-grained entities such as sentences; or alternatively coarse-grained entities such as sets of outcomes ('events'). Here I take the fine-grained approach. Indeed, I will mostly talk of probabilities attaching to sentences. One advantage, against a standard background on which logical and alethic properties attach to sentences, is that we can cleanly formulate principles that connect logic and probability without worrying about the relationship between the relata of the logical consequence relation and the bearers of probability. We can simply say, for example, that if a

[2] Supervaluationism is perhaps the leading example of a nonclassical semantics that is paired with what might be argued to be a classical logic. Supervaluational semantics allows for truth-value gaps. But, as standardly presented, across a standard propositional (or indeed quantificational) language, the associated 'global' supervaluational logic coincides with classical consequence. The issue is subtle, however. The supervaluational multiconclusion consequence relation diverges from the classical analogue. And across a minimally enriched language (including an object-language truth or definiteness operator) classical inferences rules such as conditional proof fail (Williamson, 1994, ch. 5). Supervaluational logic is a genuinely hard case to categorize (cf. Williams, 2008).

[3] For an example of this usage of 'nonclassical', and an introduction to nonclassical logics in both the narrow and the broad sense, see Priest (2001).

[4] Cf. Hájek (2012), and the chapters in this volume on *The classical interpretation and indifference principles*, *Frequentism*, *The propensity interpretation*, *Best system approaches to chance* and *Subjectivism*.

sentence is tautologous, its probability must be 1. On the other hand, given the commitment to a *subjective* interpretation of probability, the choice may seem odd. If ideal degrees of belief have to be probabilistic, it seems this requires the objects of propositional attitudes to be sentences—while believers in 'mentalese' should be happy with this, most others will not.

But there isn't a deep worry here. Suppose you hold that objects of attitudes are Fregean thoughts or Russellian structured propositions. You can then straightforwardly adapt the discussion below to your preferred setting. You already owe an account of the logic and truth-conditions of your favoured truth-bearers (and typically this can be a straightforward adaption of the usual treatment of the logic and semantics for sentences (cf. e.g. Soames (1989)). This could be classical or nonclassical. That your truth-bearers plausibly have their truth-conditions essentially doesn't prevent us from describing unintended interpretations and using them to characterize a logic in the usual model-theoretic way. The logic–probability connections appropriate to such settings will be a straightforward transcription of the sentence-based formulations below. (The real issue here is whether the *motivations* for nonclassicality extend to the propositional level. Some hold that propositional truth-conditions are broadly classical, with nonclassicality arising from the sentence–proposition relation. A case in point are treatments of reference failure which make some sentences truth-value gaps, but only because the sentences express no proposition—the propositions themselves remaining bivalent.)

One radical and minority view on the objects of belief is the 'ultra-coarse-grained' treatment of propositions argued for by Lewis (1986) and Stalnaker (1984). In one version, the propositions we take attitudes to are identified with sets of possible worlds—the possible worlds at which the sentence is true. This does motivate a rather different view of the relation between logic and probability—one which will be examined in the penultimate section of the chapter.

12.2 The Classical Framework

Consider a colour swatch, Patchy. Patchy is borderline between red and orange. The classical law of excluded middle requires that the following be true:

(LEM) Either Patchy is red, or Patchy is not red.

Many regard (LEM) as implausible for borderline cases like Patchy—intuitively there is no fact of the matter about whether Patchy is red or not, and endorsing (LEM) suggests otherwise. This motivates the development of nonclassical logic and semantics on which (LEM) is no longer a logical truth.[5] But if one doubts (LEM) for these reasons, one surely cannot regard it as a constraint of rationality that one be certain—credence 1—in it, as classical probabilism would insist. One does not have to be a convinced revisionist to feel this pressure. Even one who is (rationally) agnostic over whether or not logic should be

[5] The literature on this topic is vast. Two representatives of the contemporary debate are Wright (2001) and Smith (2008). Williamson (1994) is the most influential critic of nonclassical approaches in this area.

revised in these situations, and so has at least some inclination to doubt LEM, should not accept that non-probabilistic belief-states are irrational.[6]

We can view the distinctively classical assumptions embedded in standard probability theory in at least two perspectives. First, the standard *axiomatization* of probability (over sentences) makes explicit appeal to (classical) logical properties. Second, probabilities can be identified with convex combinations or expectations of *truth values* of sentences, where those 'truth values' are assumed to work in classical ways. We briefly review these two perspectives below in the classical setting, before outlining in the next section how these may be adapted to a nonclassical backdrop.

The following is a standard set of axioms for probability over sentences in the propositional language L:[7]

P1c. (Non-negativity) $\forall S \in L, P(S) \in \mathbb{R}_{\geq 0}$
P2c. (Normalization) For T a (classical) logical truth, $P(T) = 1$
P3c. (Additivity) $\forall R, S \in L$ with R and S (classically) inconsistent, $P(R \vee S) = P(R) + P(S)$.

Various theorems of this system articulate further relations between logic and probability:

P4c. (Zero) For F a (classical) logical falsehood, $P(F) = 0$;
P5c. (Monotonicity) If S is a (classical) logical consequence of R, $P(S) \geq P(R)$;

Believers in the so-called *Regularity* constraint on probability functions endorse yet more logical constraints on probability. They endorse the converse of (Normalization) and (Zero), saying that *only* logical truths/falsehoods take extremal probability values. I won't discuss those views further here.

(Normalization) is problematic for the logical revisionist who seeks to deny the law of excluded middle: under our interpretation of probability, it says that rational agents must be fully confident in each instance of excluded middle. But it is not the only problematic principle. Advocates of some popular nonclassical settings say that (LEM) is true, but assert the following Truth-Value Gap claim:

(TVG) It's not true that Patchy is red, and it's not true that Patchy is not red.

On this—supervaluation-style—nonclassical setting, a disjunction can be true, even though each disjunct is untrue.[8] This motivates allowing high confidence in 'either Patchy is red or Patchy isn't red', and yet ultra-low confidence in each disjunct. But this violates (Additivity).[9]

[6] The connection between logic and probability in these contexts is a major theme of Hartry Field's work in recent times. See Field (2000, 2003b,a, 2009).

[7] An alternative approach to axiomatizing probability, starting from suggestions by Popper, dispenses with the appeal to consequence, and works directly on constraints on the interaction of probability with connectives. One appealing feature of this is that one could then use probability functions so characterized as a resource for characterizing consequence. This approach has been vigorously pursued, and there are a number of extensions to nonclassical settings, such as intuitionism. See Roeper and Leblanc (1999) for a survey of both classical and nonclassical work in this tradition. The focus on 'purely logical' axiomatizations below is in a sense the dual of this approach.

[8] Some further details are given later. For supervaluations, see inter alia van Fraassen (1966); Fine (1975); Keefe (2000).

[9] Compare Field (2000). I note that sometimes, it is assumed that a 'supervaluational style' approach motivates not *low* credence but an *imprecise* credence. This is an illustration of a theme I will emphasize

Still other, dialethic, nonclassical settings allow contradictions to be true. Let L be the liar sentence ('this sentence is not true'). Some argue that the following holds:

(TC) $L \wedge \neg L$.

Advocates of this view presumably have reasonably high confidence in (TC). But (Zero) rules this out.[10]

(Monotonicity) inherits problems both from (Zero) and (Normalization). Since $A \vee \neg A$ classically follows from anything, (Monotonicity) tells us that rational confidence in excluded middle is bounded below by our highest degree of confidence in anything. And since $L \wedge \neg L$ classically entails anything, (Monotonicity) tells us that rational confidence in the conjunction of the liar and its negation is bounded above by our lowest degree of confidence in anything. But revisionists object: one such revisionist thinks we should have higher confidence that hands exist (for example) than in sentence (LEM). Another thinks we should have lower confidence in the moon being made of green cheese than we do in the conjunction of the liar and its negation.

Finally, what of (Non-negativity)? Many revisionists will find this unproblematic; notice that it doesn't appeal to (classical) logical relations explicitly at all. But the assumption that (subjective) probabilities are non-negative *real numbers* builds in, inter alia, that rational degrees of belief are *linearly ordered*. It's not crazy to question this assumption in a nonclassical setting. To take one example: MacFarlane (2010) argues that certain graphs, rather than point-like credences, capture our doxastic states in the nonclassical setting he considers.

The obvious moral from this brief review is that it would be madness for a logical revisionist to endorse *as articulating rationality constraints on belief* a probability theory that is based on the 'wrong' logic. A natural thought is to generalize the axiomatizations by switching out the appeal to classical consequence in favour of one's favoured nonclassical consequence. This is indeed the core of the approach explored below—but notice it cannot always be the whole story. The problems for (Additivity) in the supervaluational setting arise even though the relevant sentences remain inconsistent.

We turn now from axiomatics to the second perspective. This connects probabilities not to logic but directly to truth values. We presuppose an 'underlying' credence function on a maximally fine-grained partition of possibilities ('worlds'). For simplicity, we take this to be finite. The only constraints imposed on this underlying credence are that the total credences invested across all possibilities sum to 1, and that the credence in a given possibility is never less than 0. Sentences are true or false relative to these worlds. Let $|S|_w$ be a function that takes value 1 iff S is true at w, and value 0 iff S is false at w—this we call the *truth value* of the sentence at w. Each underlying division of credence c then allows us to define a function

later—that formally similar systems may allow for multiple interpretations (so 'supervaluationism' may very well pick out not a single system, but multiple such). The results below will show, however, how well the low-confidence model fits with standard supervaluational rhetoric, including the identification of truth with supertruth, and the appeal to global supervaluational consequence as the preferred consequence relation.

[10] See Priest (2006), where there is an explicit discussions of the modifications of standard probability theory required to accommodate paraconsistent logics.

from sentences to real numbers:

$$f(S) := \sum_w c(w)|S|_w$$

It turns out that such fs are exactly the probabilities over sentences. Some terminology: when a function f is a 'weighted average' of functions g_i, with weights given by coordinates $\lambda_i \geq 0$ (with the λ_i summing to 1), we say that f is a *convex combination* of the g_i. Letting the worlds w play the role of the indices i, and setting $g_w = |S|_w$ and $\lambda_w = c(w)$, the equation above meets this condition. Classical probabilities are convex combinations of classical truth values. We can also think of the probability of S so characterized as the *expectation* of S's truth value, relative to the underlying credence defined over worlds.[11] We'll call this the *convex-combination* characterization of probability.

In characterizing probability this way, the association of 1s and 0s with truth and falsity is crucial. *The True* and *the False* can't themselves be arithmetically manipulated; whereas the arithmetical manipulations of 1 and 0 make perfect sense. So why call these 'truth values' (Howson, 2008)? The answer I will explore—and extend to the nonclassical case—is that this representation is justified only because they are the degree of belief that omniscient agents should invest in S, in situations where S has that truth status. They reflect the omniscience-appropriate cognitive states; the 'cognitive loading' of the classical truth statuses. It is ultimately conventional that we represent full belief via the number 1—what we're really pointing to here is a match between the representation of truth values and the representation of maximal degree of belief.)

Since convex combinations of (classical) truth values lead to our familiar probability functions, all the problematic consequences for the logical revisionist arise once again. The revisionist faced with the convex-combination characterization of probabilities will pinpoint the appeal to *classical* truth-value distributions as what causes the trouble. Faced with classical axiomatics, the natural strategy is to consider revised principles appealing to a nonclassical consequence relation. Faced with the convex-combination characterization, the natural strategy is to explore variations where *nonclassical* truth-value distributions are appealed to.

12.3 Nonclassical Logic and Semantics

Nonclassical logic and semantics come in a wild and wonderful variety.[12] Although the results to be discussed shortly will apply to a large variety of settings, including those with (for example) infinitely many truth values, to fix ideas I set out a three-valued setting that allows us to characterize a handful of sample logics. A *Kleene truth status assignment*

[11] For those familiar with probability theory: we treat the worlds as the sample space for the probability function c, and then for any sentence S consider a random variable $t(S)$ whose value at w is equal to $|S|_w$. Where there are only finitely many worlds, the full technology of a worldly probability space isn't needed. But if there are infinitely many worlds we can still appeal to expectations of truth value, relative to the underlying credence.

[12] For a general introduction to nonclassical logics, including the Kleene logic and LP discussed below, see Priest (2001) and for further philosophical discussion, see Haack (1978).

involves, not a scattering of two statuses (Truth, Falsity) over sentences, but a scattering of three—call them for now T, F, and O. The distribution over compound sentences must accord with the (strong) Kleene truth-tables for negation, conjunction and disjunction:

A	$\neg A$
T	F
O	O
F	T

$A \wedge B$	T	O	F
T	T	O	F
O	O	O	F
F	F	F	F

$A \vee B$	T	O	F
T	T	T	T
O	T	O	O
F	T	O	F

(In the last two tables, the horizontal headers represent the truth status of A, and the vertical headers the truth status of B, and the corresponding entry the resultant truth status of the complex sentence.) We have various options for characterizing logical consequence on this basis:

Kleene logic: $A \vdash_K B$ iff on every Kleene truth status assignment, if A is T, then B is T too.

LP: $A \vdash_L B$ iff on every Kleene truth status assignment, if A is T or O, then B is T or O too.

Symmetric logic: $A \vdash_S B$ iff on every Kleene truth status assignment, if A is T, then B is T; and if A is T or O, then B is T or O.

However we characterize consequence, logical truths (tautologies) are those sentences that are logical consequences of everything; logical falsehoods are those sentences of which everything is a logical consequence; an inconsistent set is a set of sentences of which everything is a logical consequence. The strong Kleene logic is a simple example of a nonclassical logic where excluded middle is no tautology: if A has status O, then so will $\neg A$, and looking up the truth table above, so will $A \vee \neg A$. But then this provides a Kleene-logic countermodel to the claim that $A \vee \neg A$ follows from everything, since any case where B has value T and $A \vee \neg A$ value O is a countermodel to $B \vdash_K A \vee \neg A$. By contrast, excluded middle will be a tautology on the LP understanding of consequence. $A \vee \neg A$ can never have the status F; and that suffices to ensure it follows from everything in the relevant sense. LP provides us with a simple example of a paraconsistent logic—one on which explicit 'contradictions' $L \wedge \neg L$ do not 'explode'—they do not entail everything. The symmetric characterization has both features—contradictions are not inconsistent/explosive and excluded middle is no tautology.

What shall we make of these Ts, Fs, and Os? In the classical setting, we ordinarily assume that (in context) sentences are true or false *simpliciter*—that these are monadic properties that sentences (in context) either possess or fail to possess. Truth-status distributions represent possible ways in which such properties can be distributed. We could regard the Kleene distributions in the same way. The picture would then be that rather than two monadic alethic properties, there are three; but we can still ask about what the *actual* distribution is, and about the nature of the properties so distributed. Perhaps such information would motivate one choice of logic over another. A nonclassical logic motivated this way we call *semantically driven*.

But one needn't buy into this picture, to use the abstract three-valued 'distributions' to characterize the relations $\vdash_K$, $\vdash_L$ and $\vdash_S$. Hartry Field has argued for such a non-semantically-driven approach to logic in recent times. Semantics for Field does not involve *representing* real alethic statuses that sentences possess. It is rather an *instrumental*

device that allows us to characterize the relation that is of real interest: logical consequence (Field, 2009, passim). He doesn't propose that we eliminate truth-*talk* from our language—he favours a deflationarist approach to the notion—but such talk is not supposed to describe a range of 'semantic values' that sentences possess. For Field, the *T*s, *F*s, and *O*s can remain uninterpreted, since they're merely a formal tool used to describe the consequence relation. And the question of which of these categories a sentence like (LEM) falls into would simply be nonsense.

Let's suppose that we do not go Field's way, but take our nonclassical logic to be semantically driven, so that sentences have categorical properties corresponding to (one of) *T*, *F*, and *O*. What information would we like about these statuses, in order to further understand the view being put forward? Consider the classical case. Here the statuses were Truth and Falsity; and these statuses were each 'cognitively loaded': we could pinpoint the ideally appropriate attitude to adopt to each. In the case of a true sentence this was full belief (credence 1); and in the case of a false sentence, utter rejection (credence 0). We'd like to know something similar about the nonclassical statuses *T*, *F*, and *O*. If *S* has status *O*, should an omniscient agent invest confidence in *S*? If so, to what level? Would they instead suspend judgement? Or feel conflicted? Or groundlessly guess?

Call a semantics *cognitively loaded* when each alethic status that it uses is associated with an 'ideal' cognitive state. Nonclassicists endorsing a semantically-driven conception of logic may still not regard the underlying semantics as cognitively loaded. For example: Maudlin (2004) advocates a nonclassical three-valued logic (the Kleene logic, in fact) in the face of semantic paradoxes, but explicitly *denies* that there is any cognitive loading *at all* to the middle status *O*. Indeed, he thinks that the distinctive characteristic of *O* that makes it a 'truth-value gap' rather than a 'third truth value', is that it gives no guidance for belief or assertion.

The nonclassical logics we will focus on will be semantically driven, cognitively loaded, and further, will be loaded with cognitive states of a particular kind: with standard degrees of belief, represented by real numbers between 1 (full certainty) and 0 (full rejection, anti-certainty). This last qualification is yet another restriction. There's no a priori reason why the cognitive load appropriate to nonclassical statuses shouldn't take some other form—calling for some non-linear structure of degrees of belief, or suspension rather than positive partial belief, or some such. Such views motivate more radical departures from classical probabilities than the ones to be explored below.[13]

Consider the following three loads of Kleene distributions (numerical values represent the degree of belief that an omniscient agent should adopt to a sentence having that status):

Status:	T	O	F
Kleene loading:	1	0	0
LP loading:	1	1	0
Symmetric loading:	1	$\frac{1}{2}$	0

[13] Three potential examples of this are Wright's notion of a *quandary* (Wright, 2001); MacFarlane's credence profiles (MacFarlane, 2010) and whatever we should take to be the appropriate response to the partially ordered values of Weatherson (2005).

The loads differ on the attitude they prescribe for O under conditions of omniscience: utter rejection, certainty, or half-confidence respectively. They motivate informal glosses on this truth status: respectively neither true nor false; both true and false; or half-true. Furthermore, the loads correspond systematically to the logics mentioned earlier: in each case, logical consequence requires that there be no possibility of a drop in truth *value*, where the truth value is identified with the cognitive load of the truth status. (In the special case where the loads are simply 1 and 0, this corresponds to the familiar distinction between 'designated' and 'undesignated' truth statuses, and the characterization of consequence as preservation of designated status (cf. Dummett, 1959, e.g.))

We continue to use the three Kleene-based logics as worked examples. But there are many, many ways of setting up nonclassical logics. So long as the logics are semantically driven, and truth statuses are cognitively loaded with real values, then our discussion will cover them.

12.4 Probability, Truth Values and Logic

Cognitive loads give a natural way to extend the convex-combination characterization of probability.[14] Recall the classical case: for an appropriate c, the probability of each S must satisfy:

$$P(S) = \sum_w c(w)|S|_w$$

Consider the limiting case where c is zero everywhere but the actual world (i.e. conditions of 'credal omniscience') The above equation then simplifies to $P(S) = |S|_w$. Under conditions of omniscience, the subjective probability matches the numerical value assigned as S's truth value; hence, that number will be the cognitive load of the truth status. In this way, the Kleene, LP, and Symmetric loads induce three kinds of 'nonclassical probabilities', as convex combinations of the respective truth values.

A nice feature of this approach is that the axiomatic perspective generalizes in tandem with the convex-combination one.[15] Consider the following principles, for parameterized consequence relation $\vdash_x$:

P1x. (Non-negativity) $\forall S \in L, P(S) \in \mathbb{R}_{\geq 0}$
P2x. (Normalization) If $\vdash_x T$, then $P(T) = 1$
P3x. (Additivity) $\forall R, S \in L$ such that $R, S \vdash_x$, $P(R \vee S) = P(R) + P(S)$
P4x. (Zero) If $F \vdash_x$, then $P(F) = 0$;
P5x. (Monotonicity) If $R \vdash_x S$, then $P(S) \geq P(R)$;

If we pick the Kleene loads, then these five principles are satisfied by any 'nonclassical probability' (expectation of truth value), so long as we use the Kleene logic (set $x = K$). *Mutatis mutandis* for the LP and Symmetric loads and logics.

[14] See in particular Paris (2001) for this strategy. Compare also Zadeh (1968) and Smith (2010).

[15] For instances of this observation in specific settings, see e.g. Weatherson (2003); Field (2003b); Priest (2006). As we shall see, Paris (2001) gives a particularly elegant treatment of many cases.

It's useful to add two further principles—extensions and variations on (Additivity)—which are also satisfied by convex combinations of Kleene truth values:

P6x. (IncExc) $\forall R, S \in L, P(R) + P(S) = P(R \vee S) + P(R \wedge S)$
P7x. (Dual additivity) $\forall R, S \in L$, if $\vdash_x R \vee S$, then $P(R) + P(S) - 1 = P(R \wedge S)$

In the presence of (Zero) and (Normalization) respectively, plus the assumption that the conjunction of an inconsistent pair is a logical falsehood, (IncExc) will entail the original (Additivity) and (Dual Additivity). (Additivity) itself is weak in logics with few or no inconsistencies, such as *LP*; if there are no inconsistent pairs of sentences, then the antecedent is never satisfied, and the principle becomes vacuously true. (Dual Additivity) is correspondingly weak in logics with few or no tautologies, such as the Kleene logic. The weaknesses are combined in the Symmetric logic. But their generalization, (IncExc), makes no mention of the logical system in play, and so retains its strength throughout.

The connection between convex-combination and logical characterizations illustrated above is very general. Scatter truth statuses over sentences howsoever you wish, with whatever constraints on permissible distributions you like. Make sure you associate them with real-valued 'cognitive loads'—degrees of belief within $[0, 1]$, so that we can straightforwardly define the notion of possible expected truth value, by letting $|S|_w$ be equal to the cognitive load of the status that S has at w. We consider the following logic:

No drop:
$A \vdash B$ iff on every truth status assignment w, $|A|_w \leq |B|_w$.

It's straightforward to check that (P1x), (P2x), (P4x) and (P5x) will then hold of all the expected truth values.

The status of (Additivity) and its variants is more subtle. These principles make explicit mention of a particular connective, so it's no surprise that whether or not they hold depends on how those connectives behave. (IncExc) will hold iff we have the following:[16]

$$|A|_w + |B|_w = |A \vee B|_w + |A \wedge B|_w$$

Classical logic, and many nonclassical logics, satisfy this principle. But we cannot assume this holds generally.

In the classical setting, we had more than just a grabbag of principles satisfied by probabilities: we had an axiomatization complete with respect to classical expected truth values. An obvious question is whether some subset of nonclassical variants is complete with respect to nonclassical expected truth values in a similar way.

Paris (2001) delivers an elegant result on this front. Among much else of interest, he shows that the nonclassical verisons of (Normalization), (Zero), (Monotonicity) and (IncExc) deliver complete axiomatizations of a wide range of nonclassical probabilities. The conditions for this result to hold are that: (i) truth values (in our terminology: the cognitive loads of truth statuses) are taken from $\{0, 1\}$; (ii) $A \vdash_k B$ is given the 'no drop'

[16] The 'if' direction follows by the linearity of convex combinations. The 'only if' direction holds by considering the special case of probability where the underlying credence all lies on a single world, c, and hence the probability coincides with truth values.

characterization mentioned earlier; and (iii) the following pair is satisfied:

$$\begin{array}{lll} (\mathcal{T}2) & |A|_w = 1 \wedge |B|_w = 1 & \iff \quad |A \wedge B|_w = 1 \\ (\mathcal{T}3) & |A|_w = 0 \wedge |B|_w = 0 & \iff \quad |A \vee B|_w = 0. \end{array}$$

This applies, for example, to the Kleene and LP loads mentioned above, as well as the original classical case. Its application goes well beyond this: for example, to appropriate formulations of intuitionistic logic.[17]

(A side note: this is the theorem that delivers a direct extension of probability theory to many settings that are nonclassical in the 'broader' sense discussed in the introduction, for example, ones that contain modal, temporal, and conditional operators. The standard semantics for such settings will satisfy (i-iii); and the treatment of conjunction and disjunction satisfies $\mathcal{T}2, 3$.)

Beyond this, it is a matter of hard graft to see whether similar completeness results can be derived for settings that fail the Parisian conditions (one representative of which is our Symmetric logic). Drawing on the work of Gerla (2000) and Di Nola *et al.* (1999), Paris shows that a similar result holds for a finitely valued (Łukasiewicz) setting and Mundici (2006) later extended this to the continuum-valued fuzzy setting.[18]

We have already mentioned supervaluational logics. These are widely appealed to in the philosophical literature. They arise as a generalization of classical truth values, via the assumption that the world and our linguistic conventions settle, not a *single* intended classical truth-value assignment over sentences, but a set of co-intended ones. Sentences are supertrue if they are true on all the co-intended assignments, and superfalse if they are false on all of them. This allows supertruth gaps: cases where the assignments for S differ, and so it is neither supertrue nor superfalse. We shall assume that supertruth has a cognitive loading of 1, and other statuses have a loading of 0 (compare the Kleene loading earlier). The no-drop logic is then so-called global supervaluational consequence, $\vdash_s$.

This articulation of supervaluationism delivers the results mentioned earlier. For example, as a classical tautology, (LEM) is true on each classical assignment, and a fortiori true on the set of co-designated ones, so it will always be supertrue (value 1). But this is compatible with each disjunct being a supertruth gap (value 0). Invest credence in a world where this is the case, and the credences in the disjuncts can be 0 while credence in their disjunction is 1. (Additivity) and (IncExc) fail.

Axiomatizing the convex-combinations of supervaluational truth values is achieved by a theorem that Paris gives, drawing on the work of Shafer (1976) and Jaffray (1989). For the propositional language under consideration, the results show that convex combinations of such truth values are exactly the *Dempster-Shafer belief functions*. These may be

[17] For the intuitionistic case, compare Weatherson (2003). Paris reports the general result as a corollary of a theorem of Choquet (1953).

[18] The major difference between the 3-valued Kleene-based setting and the Łukasiewicz settings is the addition of a stronger conditional—and this is crucial to the proofs mentioned. It's notable that Paris provides axiomatizations not in terms of a 'no drop' logic, but in terms of the logic of 'preserving value 1'. This is possible because the 'no drop' consequence is effectively encoded in the 1-preservation setting via tautological Łukasiewicz conditionals.

axiomatized thus:

$$\begin{array}{lll} (\mathcal{DS}1) & \vdash_s A \quad \Rightarrow & P(A) = 1 \\ & A \vdash_s \quad \Rightarrow & P(A) = 0 \\ (\mathcal{DS}2) & A \vdash_s B \quad \Rightarrow & P(A) \leq P(B) \\ (\mathcal{DS}3) & P(\bigvee_{i=1}^{m} A_i) \geq \sum_S (-1)^{|S|-1} P(\bigwedge_{i \in S} A_i) & \end{array}$$

(where S ranges over a non-empty subset of $\{1, \ldots, m\}$).[19]

Completeness results of these kinds are often sensitive to the exact details of the sentences we are considering. We do *not* have a guarantee that the completeness result will generalize when we add expressive resources to the language. This is one reason why the earlier Paris result, which applies to *all* languages equipped with a semantics meeting the stated conditions, is so attractive.

($\mathcal{DS}$1-2) are simply the constraints (Normalization), (Zero), and (Monotonicity) that we met earlier. But $\mathcal{DS}3$ is something new. Sometimes called *subadditivity*, it is a new, weaker, member of the (Additivity) family. It's noticeable that what goes in place of (Additivity) is varying from setting to setting, while other principles are held constant. Why is this?

Standard axiomatizations of probability feature principles of two kinds. The first are *purely logical*: they make no mention of specific logical connectives, but put constraints on probability in terms of the logical properties of sentences. (Normalization), (Zero), and (Monotonicity) are paradigms. Axioms of the second kind are *immanent* to the nonclassical system, in that they impose constraints on sentences that involve particular logical connectives. Paradigms of this, imposing direct constraints on the distribution of probabilities over conjunctions and/or disjunctions, are (Additivity), (IncExc), (Dual Additivity), and ($\mathcal{DS}3$). Since the treatment of conjunction and disjunction can vary wildly from one nonclassical system to the next, one would not *expect* to find wholly general axiomatizations if one works with immanent axioms—one will have to indulge in case-by-case tailoring of the axioms to the particular system under investigation (or, as in Paris's result quoted earlier, impose general conditions on the semantics of the connectives that ensure that a particular immanent axiom is satisfied).

Are there purely logical axioms in the vicinity of the (Additivity) family? The following are promising. Say that Γ *x-partitions* a sentence S if the sum of truth values of the sentences in Γ always equals the truth value of S at any (nonclassical) x-world. And say that sets Γ and Δ are *x-recarvings* of one another if the sum of the truth values of sentences in Γ is always equal to the sum of the truth values of sentences in Δ.[20] With this in hand, we can formulate the following purely logical constraint on probabilities:[21]

[19] Paris's initial formulation is slightly different, and uses classical logic (p.7), but as he notes this is extensionally equivalent to the current version using the 'no drop' logic over the 'supervaluational' truth values (p.10).

[20] In the classical setting, this is a condition that Joyce (1998, 2005) calls 'isovalence'.

[21] The classical version of (Recarving) is a special case of the principle that Joyce (1998) calls 'Scott's axiom', tracing it to Scott (1964) and Kraft *et al.* (1959). The latter labeled the principle 'Generalized Additivity'.

P8x. (Partition) If Γ is a set of sentences that x-partitions S, then $P(S) = \sum_{\gamma \in \Gamma} P(\gamma)$.

P9x. (Recarving) If Γ is a set of sentences that x-recarves Δ, then $\sum_{\delta \in \Delta} P(\delta) = \sum_{\gamma \in \Gamma} P(\gamma)$.

It's easy to check that the Partition and Recarving Principles hold of all generalized probabilities.[22] Moreover (Partition) entails (Additivity) under Paris's assumptions $\mathcal{T}2$ and $\mathcal{T}3$, since R and S will then partition $R \vee S$ in the relevant sense. (Recarving) entails (IncExc) under the same assumptions, as they ensure that the set $\{R \vee S, R \wedge S\}$ recarves $\{R, S\}$.

(Partition) and (Recarving) neatly capture the *logical* structure that lies behind (Additivity), (IncExc), and the like. What additional power or generalizations one gains from the purely logical version of these axioms remains to be seen—their power depends on what partitions or recarvings are available in the particular nonclassical setting under investigation. A good test would be: can we somehow extract $\mathcal{DS}3$ in a supervaluational setting from these more abstract principles? A partial result is given in a footnote.[23]

While Paris-style completeness proofs are interesting and elegant, from a philosophical perspective, the identification of a reasonably rich body of principles that hold good of nonclassical probabilities is of philosophical interest even if we can't show them complete. Only the most radical Bayesians think that satisfying probabilistic coherence is *all* that there is to rationality; and so even if satisfying the axioms sufficed for probabilistic coherence, it would be contentious to conclude that it sufficed for rationality. On the other hand, so long as probabilistic coherence is *a* constraint on rational belief in the nonclassical setting, then what we learn from the above is that *violating* certain principles suffices for *irrationality*. A natural next question, therefore, is whether the 'nonclassical probabilities' that we have identified so far have the same claim as classical probabilities in the classical setting to provide constraints on rational belief.

[22] We argue for the recarving principle, of which the Partition Principle is a special case. Recall that an arbitrary generalized probability of any proposition is a convex combination of its truth values, say with parameters λ_w. Then $\sum_{\delta \in \Delta} P(\delta) = \sum_{\delta \in \Delta} [\sum_w \lambda_w |\delta|_w] = \sum_w \lambda_w [\sum_{\delta \in \Delta} |\delta|_w]$. By a parallel argument, $\sum_{\gamma \in \Gamma} P(\gamma) = \sum_w \lambda_w [\sum_{\gamma \in \Gamma} |\gamma|_w] =$. But by the assumption that Γ recarves Δ, we have for each w: $\sum_{\delta \in \Delta} |\delta|_w = \sum_{\delta \in \Gamma} |\gamma|_w$. So the two sums are identical, and the identity betwen the probabilities is ensured.

[23] Suppose we are working with a language with a supervaluational semantics, which includes the supervaluational 'definitely' operator D. DS is false when S is false, and also when it is neither true nor false. Otherwise it is true. First, note that since the D operator 'screens off' the nonclassical behaviour of the sentences it attaches to, we can rerun a standard classical 'inclusion-exclusion' argument from the partition principle for the special case of D-prefixed sentences, obtaining $P(\bigvee_{i=1}^m DA_i) = \sum_S (-1)^{|S|-1} P(\bigwedge_{i \in S} DA_i)$ for S a non-empty subset of $\{1, \ldots, m\}$. But it turns out in the supervaluational setting that an arbitrary conjunction $\bigwedge_{i \in S} DA_i$ is logically equivalent to the unprefixed $\bigwedge_{i \in S} A_i$. So by monotonicity twice, the RHS of the above can be written $\sum_S (-1)^{|S|-1} P(\bigwedge_{i \in S} A_i)$. On the other hand, $\bigvee_{i=1}^m DA_i$ certainly supervaluationally entails $\bigvee_{i=1}^m A_i$, so by monotonicity we have the LHS bounded above by $P(\bigvee_{i=1}^m A_i)$. Putting these together, we have $\mathcal{DS}3$.

This result holds only for a language featuring the operator D, whereas Paris's completeness result was specific to a propositional language lacking such an operator. It's possible to investigate strengthened interpretations of probabilistic constraints that bridge this gap; but for reasons of space I won't explore this here.

12.5 Foundational Considerations

In many well-behaved nonclassical settings, we have seen a nice generalization of probability theory in prospect. But is this just coincidence? Or can we argue that this is the *right* way to theorize about subjective probabilities in such a setting?

I will focus on two arguments for 'probabilism' familiar from the classical case: the Dutch-bookability of credences that violate the axioms of probability, and the accuracy-domination arguments advocated by Joyce (1998, 2009). Jeff Paris has shown how the first can be generalized, showing (Paris, 2001) that credences that are not *convex combinations of truth value* in the relevant sense are Dutch-bookable, and conversely. And for similar formal reasons, such credences are also 'accuracy-dominated' (a converse is sometimes available).[24]

In the classical case, accuracy-domination arguments consist in taking a belief state b, and assessing it at each world w for its degree of 'accuracy'. How accuracy should be measured is the leading issue for this approach; but in all cases, the starting point is to compare a degree of belief (within $[0, 1]$) to the 'truth value' of the sentence in question. But comparing a number with *Truth* or *Falsity* is not terribly tractable. So one standardly compares a given degree of belief with the cognitive loading of the truth status—how close, overall, the degrees of belief are to the 1s and 0s that an omniscient (perfectly accurate) agent would have in that world.

The result (relative to many plausible ways of measuring accuracy) is the following: if one's beliefs b are not (classical) probabilities, then it is always possible to construct a rival probabilistic belief state c such that c is more accurate than b *no matter which world is actual*. If such accuracy-domination is an epistemic flaw, then only probabilistic belief states can be flawless. This is offered as a rationale for why subjective *probabilities* in particular should be constraints on ideally rational belief.

What's important for our purposes is that the argument generalizes.[25] As earlier, suppose the cognitive loading of some nonclassical truth statuses are real numbers between $[0, 1]$. We use the very same accuracy measures as previously, to measure closeness of beliefs to these nonclassical truth values. And it turns out that if the belief state is not representable as a convex combination of truth values then it will be accuracy-dominated. The accuracy-based arguments for probabilism thus offer a justification for the claim that nonclassical probabilities, as characterized in the previous section, should indeed play the role of constraints on rational partial belief. (Of course, whether it's a good justification in either setting is contested. See Hájek (2008a).)

[24] See de Finetti (1974) for the formal background to both results (in the latter case, with a quite different interpretation of its significance). Williams (2012a) examines the relation between the two results—it is essentially identical for the leading 'Brier Score' explication of accuracy. In that setting, and in others where accuracy is explicated by what are known as 'proper scoring rules', a converse to the accuracy-domination result is available; no probability will be accuracy-dominated. For discussion of the philosophical significance of converses to accuracy-domination arguments, see Williams (2012b). A rather different connection between Dutch book foundations for probability and nonclassical (intuitionistic) logic is argued for in Harman (1983).

[25] At least, the results in Joyce (1998) generalize. As discussed in Williams (2012b) the situation is much more complex for the main argument in Joyce (2009). The variation in proofs is significant, since different assumptions about the accuracy measure are involved in each case.

Perhaps the most familiar foundational justification for the claim that rational partial beliefs must be probabilistic comes from Dutch book arguments. The key claim is that one's degrees of belief are fair betting odds in the following sense: if offered a bet that pays £1 if A, and £0 if $\neg A$, then if you believe A to degree k, you should be prepared to buy or sell the bet for k. Suppose that degrees of belief do play this role. Then if b is an improbabilistic belief state, there is a set of bets—a 'Dutch book'—such that you are prepared to buy each bet within it, but which ends up giving you a loss *no matter what*. Pragmatically viewed, a set of beliefs that open you up to sure losses may seem flawed. Alternatively, one might think that the belief state is flawed because it commits you to viewing the book as both to be accepted (because consisting of bets that your belief state makes you prepared to accept) and also obviously to be rejected (since it leads to sure loss). So you are committed to inconsistency.

Dutch book justifications for probabilism, like accuracy-domination arguments, are controversial (see Hájek (2008a,b) for a review and extension of criticisms). But independently of whether they persuade, are they adaptable to our case?

Suppose one has bought a bet that pays out 1 if A and 0 otherwise. If one is in a nonclassical setting, one can be faced with a situation where A takes some nonclassical truth status. The return on such a bet then depends on how the bookie reacts. Call a real number $k \in [0, 1]$ the *pragmatic loading* of a truth status X, just in case the right way for the bookie to resolve such a bet, given that A has status X, is to give the gambler £k.[26] Clearly the pragmatic loading of classical truth should be 1, and the pragmatic loading of classical falsehood is 0. Just as with cognitive loads of nonclassical truth statuses, there are many many ways one might consider assigning pragmatic loads (and just as with cognitive loads, there are pragmatic loads for truth statuses that don't fit the above description—the option of 'cancelling the bet' for example—as well as the option to deny that truth statuses *have* any identifiable pragmatic loading).

Suppose we have real-valued pragmatic loads for truth statuses, however. Then we can make sense of resolving bets in a nonclassical setting, and can consider what kinds of belief states are immune to Dutch books. Happily, the answer is just as you would expect: immunity to Dutch books is secured when (and only when) the belief state is a 'nonclassical probability'—a convex combination of the relevant truth values.[27]

It's worth noting that in this last result, the 'truth value' of a sentence refers to the *pragmatic loading* of the relevant truth status, whereas in the previous results it referred to the *cognitive loading* of the truth statuses. If they differed, then we might have inconsistent demands—for example, if the cognitive loading of the 'other' status was 0.5 (omniscient agents are half-confident in A, when A is *O*), but its pragmatic loading was zero (one doesn't receive any reward for a bet on A, given that it is half-true) then being 0.5 confident in $A \wedge \neg A$ might be entirely permissible from the accuracy-domination point of view, but still make you Dutch-bookable. The way to avoid this, of course, is to have cognitive and pragmatic loads coincide. It is interesting to speculate on *whether* they should coincide,

[26] Compare the stipulations in Paris (2001) on the returns of bets in a nonclassical setting. On this description there's room for a kind of meta-uncertainty about what the pragmatic loading is—which could be modelled by allowing a wider class of 'truth-value' distributions over worlds corresponding to all the possible pragmatic loading distributions the agent is open to.

[27] The result follows from Dutch book arguments for expectations in de Finetti (1974) and is interpreted in the way just mentioned in Paris (2001). For more discussion, see Williams (2012a).

and if so *why*. I can imagine philosophers taking cognitive value as primary, and arguing on this basis that the *right* way to resolve bets accords with the pragmatic loading; but I can equally envisage philosophers arguing that pragmatic loads are primary, and that these *give the reasons* why a particular cognitive loading attaches to a truth status. I can also imagine someone who takes both as coprimitive, but argued ('transcendentally') that they must coincide, otherwise rationality would place inconsistent demands on agents.

Both Dutch book and accuracy arguments—and much of the debate between their advocates and critics—can be replayed in a nonclassical setting. This should bolster our confidence that we have the *right generalization* of probability theory for the cases under study. And none of the results just mentioned make *any* assumptions about the particular kinds of truth-value distributions or logical behaviour of the connectives in question—other that the truth values, in the relevant sense, lie within $[0, 1]$. These are extremely general results.

12.6 Conditional Probabilities and Updating

Subjective probability without a notion of conditional probability would be hamstrung. If we are convinced (at least pro tem) that we have a nonclassical generalization of probability, then the immediate question is how to develop the theory of conditional probability within this setting. Three approaches suggest themselves. The first is simply to carry over standard characterizations of conditional probabilities, the ratio formula (restricted to cases where $P(A) \neq 0$; I often leave such constraints tacit in what follows):

$$P(B|A) := P(B \wedge A)/P(A)$$

The second is to investigate axiomatizations of probability in which conditional probability is the basic notion (of course, if left unchanged, these lead to classical probabilities). One investigates variations of these axioms, much as we did for the unconditional case above (compare Roeper and Leblanc (1999)). Thirdly, we can look to the work we want conditional probability to do, and try to figure out what quantity is suited, within the nonclassical setting, to play that role. It is this third approach we adopt here, with a focus initially on the role of conditional probability in *updating* credences.

Conditional probabilities will be two-place functions from pairs of propositions to real numbers, written $P(\cdot|\cdot)$. The key idea will be that this should characterize an update policy: when one receives total information A, one's updated unconditional beliefs should match the old beliefs conditional on A: $P_{\text{new}}(\cdot) = P_{\text{old}}(\cdot|A)$. If updating on information isn't to lead us into irrationality, then a minimal constraint on conditional probabilities fit to play this role is that the result of 'conditioning on A', as above, should be a probability. (It turns out, incidentally, that straightforwardly transferring the 'ratio formula' treatment of conditional probabilities can violate this constraint.)[28]

[28] Suppose that we are working within the 'symmetric/half truth' nonclassical setting, suppose $P(A) = P(\neg A) = P(A \wedge \neg A) = 0.5$—which is certainly permitted by the relevant nonclassical probabilities. Now

Classical conditionalization on A can be thought of as the following operation: one first sets the credence in all $\neg A$ worlds to zero, leaving the credence in A-worlds untouched. This, however, won't give you something that's genuinely a probability (for example, the 'base credences' no longer sum to 1). So one renormalizes the credences to ensure we do have a probability, by dividing each by the total remaining credence $P(A)$.

We could generalize this in several ways, but here is the one we will consider. Take the first step in the classical case: we wipe out credence in worlds where the proposition is false (truth value 0) and leave alone credence in worlds where the proposition is true (truth value 1). Another way to put this is that the updated credence in w, $c_A(w)$ (prior to renormalization) is given by $c(w)|A|_w$: the result of multiplying the prior credence in w by the truth value of A at that possibility. Since we have real-valued truth values in our nonclassical settings, we simply transfer this across. The credence is scaled in proportion to *how true A* is at a given possibility. Renormalizing is achieved just as in the classical setting, by dividing by the prior credence in A.[29] Notice that by focusing on how the underlying credence c is altered under conditionalization, we have guaranteed that the function $P_A(X)$ defined by this procedure will be a convex combination of truth values, and so a nonclassical probability in our sense. We set $P(X|A) := P_A(X)$.

The characterization of the update procedure can be set down as follows:

$$P_A(X) = \sum_{w \in W} \frac{c_A(w)}{P(A)} |X|_w$$

Expanding the right-hand side, this gives the following fix on conditional probability:

$$P(X|A) = \frac{\sum_{w \in W} c(w)|A|_w|X|_w}{P(A)}$$

Now, if we have a connective $\circ$ such that for arbitrary A and B, at any w:

$$|A|_w|B|_w = |A \circ B|_w$$

then it follows:

$$P(X|A) = \frac{\sum_{u \in W} c(u)|A \circ X|_u}{P(A)} = \frac{P(X \circ A)}{P(A)}$$

Certain nonclassical settings already contain the connective $\circ$—in the classical, supervaluational, Kleene and LP settings, $\wedge$ plays this role, and so the familiar ratio formula with relatively familiar conjunctive connectives is derived. A more exciting example is the main conjunctive connective of the product fuzzy logic (cf. Hájek, 1998) In other settings, it is a well-defined truth function, but would require an extension of the language to

consider $P(A \wedge \neg A|A)$. By the ratio formula, this would be $P(A \wedge \neg A \wedge A)/P(A) = P(A \wedge \neg A)/P(A) = 0.5/0.5 = 1$. So $P_{\text{new}}(A \wedge \neg A) = 1$. But no probability (convex combination of truth values) in this setting can have this exceed 0.5.

[29] To see this, note that the result of the process is to give a 'base credence' over worlds which may add to less than 1. The sum total is given by $\sum_{w \in W} c_A(w) = \sum_{w \in W} c(w)|A|_w$. But by construction this is exactly $P(A)$. Hence dividing by $P(A)$ will renormalize the base credence, making it sum to 1, after the procedure described above.

introduce. But it is not automatic that the conditions for $\circ$ can be met by a truth function, in arbitrary nonclassical systems. Consider, for example, the Symmetric loading of the Kleene assignments. A sentence that has the truth status *O* gets the truth value 0.5. So $A \circ A$ would by construction have to take the truth value 0.25; $(A \circ A) \circ A$ would have to take the value 0.125, and so on. But in the symmetric setting, there are no truth statuses that have these loads. (A fortiori, $\circ$ is clearly *not* the Symmetric connective $\wedge$.) The process of conditionalization that was described works perfectly well in the Symmetric setting as a way to shift from one nonclassical probability to another on receipt of the information that *A*. It's just that it doesn't have a neat formulation that mirrors the ratio formula.[30]

Much more on these nonclassical conditional probabilities in the fuzzy logic setting is available in Milne (2007, 2009)—who cites Zadeh (1968) as the source for the conception. Milne shows how to provide a synchronic Dutch book argument for this characterization of conditional probability, relative to the assumption (i) that conditional probabilities give fair betting odds for conditional bets; (ii) that nonclassical conditional bets are to be resolved a certain way (in particular, that they are 'progressively more and more called off' as the truth value of the condition gets lower and lower).[31] As Milne emphasizes, the assumption (ii) is crucial; in principle, there are many ways one might consider handling conditional bets in this setting, which would vindicate different conceptions of conditional probability. Williams (2012a) gives a nonclassical generalization of the Teller-Lewis diachronic Dutch book argument, but (although it gives us non-trivial information) it is even worse at giving leverage on the crucial case of conditionalizing on sentences that at some worlds take nonclassical values.

The real test for a proposed generalization of conditional probabilities lies in its applications—as an update procedure and elsewhere. To give a flavour of some important ways in which it generalizes classical conditional probability, we show how some key results generalize.

[30] Suppose that we have another connective $A \oplus B$ that is dual to $\circ$ in the following sense: $|A \oplus B|_w = |A|_w + |B|_w - |A \circ B|_w$. Then, by an earlier note, (IncExc) will hold for probabilities involving $\circ$ and $\oplus$ as the conjunction and disjunction symbols. It's worth noting that even if a setting has the $\circ$ connective, it needn't have its dual. In supervaluational logic, $\wedge$ coincides with $\circ$, but $\vee$ does not coincide with $\oplus$—as we can see by noting that $A \oplus \neg A$, unlike the supervaluational $\vee$, always takes truth value zero when both *A* and its negation have truth value zero. An alternative way to introduce a disjunction via $\circ$ is through the De Morgan identity: $|A \vee B|_w = |\neg(\neg A \circ \neg B)|_w$. In a supervaluational setting, this will be the normal supervaluational disjunction, for which we already know that (IncExc) does not hold. The general moral is that one can introduce an (IncExc) supporting dual of $\circ$, but there's no guarantee that it exists in a given setting; and one can find a corresponding notion of disjunction that is definable in any system with $\circ$ and $\neg$, but there's no guarantee that it is dual to $\circ$ in the way that (IncExc) requires.

[31] With classical presuppositions, the assumption is that a bet on *A* conditional on *C* with prize 1 and price β will return the prize if *AC* is the case; will return nothing if $A\bar{C}$ is the case; and the bet is called off (with a return of the initial stake) if the condition *C* is false. In the general case, we need to consider what happens to the bet in situations when *C* is partially true. Here is the stipulation: a conditional bet on *A* given *C* at prize 1, price β will have part of the stake returned, and the potential prize decreased, in proportion to the falsity of *C*. Modulo this, the returns depend on *A*'s truth value as in the categorical case. The overall return of the unit bet above is therefore $|C|_w(|A|_w - \beta)$ at w. The philosophical premise we need is that the fair price for a conditional bet so construed is exactly the conditional probability of *A* on *C*.

The analogue of Bayes' theorem is immediate:

$$P(A|B) = \frac{P(A \circ B)}{P(B)} = \frac{P(A \circ B)}{P(A)} \frac{P(A)}{P(B)} = P(B|A) \frac{P(A)}{P(B)}$$

Further key classical results also carry over:

1. **Lemma.** Assume that $\forall w, |\neg A|_w = |1 - A|_w$. Then $P(C) = P(C \circ A) + P(C \circ \neg A)$.
   Proof. First note that relative to arbitrary w,

   $$|C| = |C|(|A| + 1 - |A|) = |C|(|A| + |\neg A|) = |C||A| + |C||\neg A| = |C \circ A| + |C \circ \neg A|.$$

   For arbitrary nonclassical probability P, there's an underlying credence-over-worlds c such that $P(A) = \sum_w c(w)|A|_w$. So in particular

   $$P(C) = \sum_w c(w)|C|_w = \sum_w c(w)(|C \circ A| + |C \circ \neg A|).$$

   But this in turn is equal to:

   $$\sum_w c(w)|C \circ A| + \sum_w c(w)|C \circ \neg A| = P(C \circ A) + P(C \circ \neg A)$$

   as required.
2. **Corollary.** $P(C) = P(C|A)P(A) + P(C|\neg A)P(\neg A)$. Follows immediately from the above by the $\circ$-ratio formula for conditional probability.

Recall that Γ is a *nonclassical partition* if in each world, the sum of the truth values of the propositions in Γ is 1 (thus our assumption that $|\neg A| = 1 - |A|$ ensured that $A, \neg A$ was a partition). Then replicating the above proof delivers:

1. **Generalized Lemma.** $P(C) = \sum_{\gamma \in \Gamma} P(C \circ \gamma)$, so long as Γ is a partition.
2. **Generalized Corollary.** $P(C) = \sum_{\gamma \in \Gamma} P(C|\gamma)P(\gamma)$, so long as Γ is a partition.

It's nice to have this general form since there is some settings (supervaluational semantics for example) where the truth values of A and $\neg A$ don't sum to 1; the partition-form is still applicable even though the first result is not.

Another useful result is that, if P_C is the probability that arises from P by updating on C, then $P_C(A|B) = P(A|B \circ C)$. This follows straightforwardly from the ratio formula. For:

$$P_C(A|B) = P_C(A \circ B)/P_C(B) = \frac{\frac{P(A \circ B \circ C)}{P(C)}}{\frac{P(B \circ C)}{P(C)}} = \frac{P(A \circ B \circ C)}{P(B \circ C)} = P(A|B \circ C)$$

As an application, we can use the above results together with the basic convex-combination characterization of nonclassical probabilities to derive the following two 'expert principles':[32]

$$P(S|t(S) = x) = x$$

$$P(S) = \sum_x x \cdot P(t(S) = x)$$

For the first, the characterization of conditional probabilities, and expansion by the convex-combination characterization and the definition of $\circ$ gives:

$$P(S|t(S) = x) = \frac{P(S \circ t(S) = x)}{P(t(S) = x)} = \frac{\sum_w c(w)|S \circ t(S) = x|_w}{\sum_w c(w)|t(S) = x|_w} = \frac{\sum_w c(w)|S|_w|t(S) = x|_w}{\sum_w c(w)|t(S) = x|_w}$$

But $|t(S) = x|_w$ takes value 1 at worlds where the truth value of S is x, and 0 otherwise. So we can rewrite the above summing over those worlds where its truth value is indeed x. And of course, at all those worlds, $|S|_w = x$, from which we get our result:

$$= \frac{\sum_{w:|S|_w=x} c(w)|S|_w}{\sum_{w:|S|_w=x} c(w)} = \frac{\sum_{w:|S|_w=x} c(w)|S|_w}{\sum_{w:|S|_w=x} c(w)} = \frac{x \cdot \sum_{w:|S|_w=x} c(w)}{\sum_{w:|S|_w=x} c(w)} = x$$

The second of the two expert principles follows from the first by the generalized corollary above, applied to the partition given by all sentences of the form $t(S) = x$. The net result is a derivation of the 'expert' principle that, for rational believers, one's degree of belief in S should match one's expectation of its truth value. (In an earlier slogan, we used a similar 'expectational' gloss on nonclassical probabilities—but that required appeal to the underlying credence distribution over worlds. The result here is formulated purely in terms of the sentential probability P, and matches one standard classical way of calculating the expected value of a random variable t.)

This all looks promising. On the other hand, there are some surprising divergences from how we might expect a conditional probability to behave. Conditional probability so generalized does not guarantee that $P(A|A) = 1$. Consider the Symmetric-loaded Kleene setting. Suppose all credence is invested in a world w in which A has truth value 0.5. It turns out by the recipe above that $P(A|A) = 0.5$.[33] This is surprising! Here is an application: probabilistic independence is standardly defined as holding between A and B when $P(A|B) = P(A)$. But in the case just mentioned, $P(A|A) = 0.5 = P(A)$, so A will be probabilistically independent of itself. This shows that the epistemic significance of independence (so defined) and the use of conditional probabilities in confirmation theory more generally will need to be looked at carefully.[34]

[32] Compare Reflection (van Fraassen, 1984) and the Principal Principle (Lewis, 1980). Recall that $|t(S) = x|_w = 1$ iff $|S|_w = x$; otherwise it takes value 0. The second displayed summation makes sense if x takes finitely many values in $[0, 1]$; otherwise we will have to switch to integral formulations.

[33] Multiplying the underlying credence by the truth value of A in any world other than w gives 0; at w it gives 0.5. Renormalizing takes this back up to 1—the underlying credence distribution is unchanged, so $P(A|A) = P_A(A) = P(A) = 0.5$.

[34] Thanks to Al Hájek for this example. Note, however, that even in the classical case, A can be probabilistically independent of itself if it has probability 1. Even in the classical case, this is a little

12.7 Jeffrey-Style Decision Theory

An important application of probability is within the theory of rational decision-making. We want to say something about a decision situation taking the following form: there is a range of actions A. There are factors $S \in \Gamma$, which fix the consequences of the action. Γ form a partition, and we are uncertain which element of that partition obtains. We are in a position to judge the desirability of the total course of events, representable by $A \wedge S$. But our uncertainty over which S obtains means that we have work to do to in order to figure out the desirability of A itself.

Jeffrey's decision theory (Jeffrey, 1983) provides a way to calculate the desirability D of a course of action, from the desirability of the outcomes, plus one's subjective probabilities. The desirability of the action is a weighted average of the desirability of the outcomes, with the weights provided by how likely the outcome is to obtain, given you take the action. The recipe is:

$$D(A) = \sum_{S \in \Gamma} P(S|A) D(A \wedge S)$$

Notice the crucial role given to conditional probabilities.

The application of probabilities within the theory of decision-making is important, and if we couldn't recover a sensible account, this would render the whole enterprise of nonclassical (subjective) probability less interesting. As a proof of principle, I'll show that Jeffrey's recipe can indeed be generalized. I do not claim here to *justify* this as the *right* theory of decision in the nonclassical setting, but just to show that such theories are *available*.

Here's one way in which the Jeffrey decision rule can arise. Start by introducing a valuation function from worlds to reals, v—intuitively, a measure of how much we'd like the world in question to obtain. Then the desirability of an arbitrary *sentence* A is defined as follows:

$$D(A) := \sum_{w} P(w|A) v(w)$$

Now, one might wonder if this is the *right* way to define desirability; but there is no question that it is *well-defined* in terms of the specific underlying valuation v. Now take any partition Γ of sentences (in the generalized sense of partition of the previous section). By the corollary in that section, applied to the nonclassical probability P_A that arises from conditioning on A, we have:

$$P(w|A) = P_A(w) = \sum_{S \in \Gamma} P_A(S) P_A(w|S)$$

Using another fact noted there, $P_A(w|S) = P(w|S \circ A)$, and putting these two together and substituting for $P(w|A)$ in the definition above, we obtain:

$$D(A) = \sum_{w} [\sum_{S \in \Gamma} P(S|A) P(w|S \circ A)] v(w)$$

strange—what we see here is that in the nonclassical case the phenomenon cannot be confined to cases where A initially has an extremal value.

Rearranging gives:

$$D(A) = \sum_{S \in \Gamma} P(S|A)[\sum_{w} P(w|S \circ A)v(w)]$$

But the embedded sum here is by construction equal to $D(S \circ A)$. Thus we have:

$$D(A) = \sum_{S \in \Gamma} P(S|A)D(S \circ A)$$

This is the exact analogue of the Jeffrey rule. So valuations over worlds allow us to define a notion of desirability that satisfies the generalized form of Jeffrey's equation.

What of the axioms for Jeffrey-Bolker decision theory? I won't investigate these here, but I will note that at least some require modification. For example, the 'averaging' axiom of Jeffrey-Bolker decision theory tells us that if A and B are inconsistent, then if $A \succ B$, then $A \succ (A \vee B) \succ B$. But on a supervaluational model of decision theoretic utility constructed as above, this can fail. Consider a situation where the value assigned to w and u is 1 in each case, and where $|A|_w = |\neg A|_w = 0$ and $|A|_u = 1, |\neg A|_u = 0$. In a supervaluational setting, $|A \vee \neg A|$ is 1 at any world. Then $D(A) = 1, D(\neg A) = 0$ but $D(A \vee \neg A) = 2$, a violation of the axiom. The underlying issue, I suspect, is that the axiom is written on the assumption that when A and B are mutually exclusive, they form a partition of $A \vee B$—which is perfectly true in a classical setting, but fails in the supervaluational setting. What can be read off the definitions above is that whenever (i) C is such that A and B form a partition of it (i.e. $\forall w, |C|_w = |A|_w + |B|_w$), and (ii) $A \circ C \succ B \circ C$, then we have $A \circ C \succ C \succ B \circ C$. This has the standard axiom as a special case, once classical assumptions are added.

12.8 Alternative Approaches

Having explored the interaction of nonclassicality and probability under the understanding of those notions identified in section 1, this section takes a step back and considers briefly how things might look if we varied our starting assumptions.

The first variation we will consider concerns the items to which probabilities attach. We have been talking as if both probabilities and logical properties attach to sentences. We emphasized earlier that the focus on linguistic bearers is inessential—all the above could be transferred to probabilities and logic pertaining to Fregean thoughts, Russellian structures, and similar fine-grained 'propositions'. However, a very common approach to probability theory takes the objects of probabilities to be *coarse-grained*. For example, one finds a probability defined in terms of a triple (Ω, F, P), where Ω is a set (the 'sample space'), the event space F is an algebra of subsets of Ω, and P is a function from F to real numbers. We then have the familiar Kolmogorov axioms:

P1. (Non-negativity) $\forall E \in F, P(E) \in \mathbb{R}_{\geq 0}$
P2. (Normalization) $P(\Omega) = 1$
P3. (Additivity) $\forall D, E \in F$ with D and E disjoint, $P(D \cup E) = P(D) + P(E)$

On one reading, Ω could be the set of possible worlds, and then F would be coarse-grained *propositions* in the sense of Lewis (1986) and Stalnaker (1984): sets of possible worlds. Lewis and Stalnaker argue that coarse-grained entities such as these are the objects of attitudes—a radical and minority position that means that even ordinary agents cannot take different attitudes to necessarily equivalent propositions. (An intermediate position here is to allow fine-grained entities as the objects of attitudes, so that different attitudes to necessarily equivalent claims are possible, but insist that *ideal* agents' degrees of belief across mentalese sentences should be such that a probability function is induced across the coarse-grained propositions such sentences express.) The interesting thing about this setting from our perspective is that *logic* seems to have disappeared from view. To be sure, we have analogues of 'inconsistency' (disjointness)—but the classicality is not explicit.

A natural diagnosis is that classical logic is tacitly built into this framework—by the assumption that the event space F forms a Boolean algebra. *Quantum* probabilities are developed exactly on this diagnosis on the (highly contentious) quantum logic approaches to quantum mechanics. Quantum events are held to form a non-distributive lattice, corresponding to the structure of subspaces of Hilbert space, rather than to the structure of a subsets of a classical sample space (Hughes, 1992, chs. 8, 9). Probabilities attaching to these quantum events will be distinctively 'nonclassical'.[35] More generally, to generate nonclassical probabilities we consider all sorts of alternative algebraic structures for the event space, and study analogues of the standard probability axioms in that setting. Algebraic logic provides a rich set of resources for those interested in pursuing this approach (Jansana, 2011). The variation of the algebra can be radical (replacing Boolean by Heyting algebras, for example) or it could be more minor (minor variations in the standard closure conditions of the algebra, for example).

In the discussions in previous sections, at various points I assumed an underlying space of 'worlds' at which sentences take truth values, one of which is 'actual', and determines a definite truth-value distribution. The final formulations are often statable without making this assumption (hence those formulations are available as a piece of formalism even to one who rejects the above), but it was freely used in their motivation and justification. It's worth bearing in mind that some of the more radical motivations for nonclassicism might make this assumption questionable. For example, the advocates of quantum logic interpret the Kochen-Specker theorems as ruling out this kind of picture; a Dummettian anti-realist about the past might find it hard to accommodate; and in his work developing a nonclassical logic to evade the liar paradox, Field (2008) specifically argues against the idea there is a definite distribution of actual (even nonclassical) truth values. Perhaps the algebraic approach, together with a philosophical interpretation of what the items in the 'event space' are to be, will seem to these theorists a more attractive philosophical foundation than the one sketched here.

[35] The Hilbert space structure and the numbers assigned are common ground in quantum mechanics; but other approaches will offer alternative interpretations of the formalism—for example, identifying some Boolean subalgebra as the 'real' event space, with standard classical probabilities defined across it, with numbers assigned to points outside this subalgebra given an alternative physical interpretation.

We've just considered alternative approaches to our question that vary the objects of probabilities. Another alternative is to vary what 'probability' is to mean. Now, some interpretations of probability are tightly connected with epistemic issues in a way that allows much of the above discussion to go through. Evidential probabilities, as in Williamson (2000) are a case in point. Prima facie we expect that our evidence can discriminate between necessarily equivalent propositions. I may have strong evidence for the liquid in my glass being water, without having evidence that it is H_2O. This makes a fine-grained setting natural (though a coarse-grained theorist may respond instead by allowing 'possible worlds' into the sample space that are not *metaphysically* possible). The nonclassicist who rejects Patchy's being red or not red certainly seems to need a treatment on which this does not get assigned evidential probability 1. There's still some plausibility in the claim that assignments of evidential probability can't be accuracy-dominated; and that setting betting odds according to the evidence shouldn't lead to a sure loss. So while the material certainly needs to be examined carefully and reworked under the new interpretation, it still seems directly relevant.

By contrast, under an interpretation of probability as objective chance, the parameters of the discussion are changed more radically. First, one has to decide what the vehicles of chance are. Unlike belief, there's little motivation to think that necessarily equivalent propositions should be able to take different probabilities. That makes the coarse-grained formulation of probability, on which chances attach to sets of possible worlds, a natural one. This takes us back to the issues discussed earlier in this setting, whereby nonclassicality has to impact, if at all, in the algebraic structure of event space. On the other hand, we may still want to make sense of thought and talk about the objective chance of vaguely-specified events. In ordinary life, we might want to know about the chance of a ball in a bag's being red. In special sciences, chances may attach to macro-events *e*, such that some physically possible course of micro-events will make it vague whether *e* obtains. The vagueness blocks a straightforward translation into the obvious coarse-grained setting. Likewise, theses that connect chance to subjective belief, such as the Principal Principle, are often stated in a way that presupposes a common domain of entities to which both degrees of belief and objective chances attach. I take the moral of this to be that *even if* the underlying chancy phenomena are best thought of as pertaining to coarse-grained events or propositions, we may owe an account of an induced chance-function across fine-grained propositions or sentences. At this point the main thread of discussion in this chapter is again relevant. The nonclassicist who rejects the claim that Patchy is red or not red certainly won't want to say that the objective chance of Patchy's being red or not red is 1. So if this chance-attribution is well-formed at all, it seems to require nonclassical objective chances. Of course, the earlier motivations for the particular formulations of nonclassical probabilities (such as the Dutch book and accuracy arguments) presupposed the subjective interpretation. We might hope nevertheless to motivate nonclassical probabilities as formulations of objective chances by way of the Principal Principle, plus nonclassical probabilism about degrees of belief; the nonclassical analogue of the project carried out in Lewis (1980).

There are plenty of other reinterpretations of probability to consider. In each case, the nonclassical probabilities we've been looking at are an obvious resource, but as has been illustrated, the details matter.

12.9 Conclusions

To recap the main points of our discussion:

A. Nonclassical probabilities can be viewed as convex combinations of nonclassical truth values, and standard principles of probability can often be carried over to the nonclassical case if we substitute an appropriate nonclassical logic for the classical one. The appropriate nonclassical logical entailment can be generally characterized as one guaranteeing no drop in truth value. In Section 6, we showed that truth values behave as 'experts' relative to these probability functions.
B. The truth values concerned should be thought of as the cognitive loads of the nonclassical truth statuses. But the general recipe may break down if (a) one does not endorse a semantically-driven conception of logic (one will not have 'truth statuses' to play with); (b) one does not regard the statuses as cognitive loaded; or (c) the cognitively loads are not representable as real-valued degrees of belief. As discussed in Section 7, this approach is also undermined if one insists on defining probabilities not over fine-grained truth-bearers (whether sentences, thoughts or structured propositions) but over coarse-grained, algebraically structured events or truth conditions.
C. The nonclassical probabilities so defined can be justified as constraints on rational belief via analogues of Dutch book and accuracy-domination arguments.
D. A notion of conditional probability can be characterized, that preserves important features of classical conditional probability. It satisfies a ratio formula, using a connective that is available in some but not all nonclassical settings.
E. An analogue of Jeffrey's recipe for calculating desirability of actions with respect to an arbitrary partition of states is available.
F. Though the discussion is conducted in the context of a subjective interpretation of probability, the theory of nonclassical probabilities that emerges is a resource for probability theory under other interpretations, though each application raises new philosophical issues.

This provides a rich field for further investigation.

A. Studying *axiomatizations* of nonclassical probabilities is an open-ended task. Can we extend the results of Paris, Mundici, et al., and get a more general sense of what set of axioms is sufficient to characterize convex combinations of truth values? One key step was outlined earlier: to switch from 'immanent' axiomatizations (such as Additivity) to purely logical ones (such as Recarving).
B. We have focused on cases where 'truth values' (the cognitive loading of nonclassical truth statuses) take a particularly tractable form: represented by reals in $[0, 1]$. Can we get a notion of nonclassical probability in a more general setting, where the cognitive loads are not linear ordered, or where some truth statuses are missing such loading altogether? Perhaps the notion of expectations of non-real-valued random variables may provide a lead here.

C. What is the relation between the cognitive loading of a nonclassical truth status (appealed to directly in the accuracy-domination argument) and the pragmatic loading (relevant to the generalized Dutch book argument). Must they coincide? If so, why?
D. How much of the theory of conditional probability transfers to the nonclassical setting? Can confirmation theory, in particular, be preserved in the nonclassical setting?
E. Many foundational and formal questions about nonclassical decision theory deserve exploration. Are analogues of classical representation theorems available, relative to a set of rational constraints on qualitative preference? What are the appropriate generalizations of the qualitative constraints of Jeffrey-Bolker theory? And can other forms of decision theory find expression in the nonclassical setting?
G. There are many possible interpretations of probability for which we could raise questions analogous to those discussed here. The chance and evidential interpretations have only been briefly discussed—and there's plenty more to say about these, and the relation between different forms of probability in the nonclassical setting. One particularly interesting interpretation to study is the *logical* intepretations of probability, on which the aim is to articulate the degree to which one proposition supports another, as a putative generalization of the total support captured by ordinary logical consequence. Surely *this* should interact directly with a nonclassical logic.

It's worth emphasizing a remark we made right at the start. It is not only convinced revisionists who need to be concerned about these issues. Anyone not dogmatically opposed to logical revision needs to take interest. For, prima facie, if one is open to the possibility that excluded middle fails in some cases, for example, then one shouldn't invest full confidence in each of its instances. And yet, it does not seem that one is irrational in harbouring such doubts, as the interpretation of classical probabilities as constraints on rational belief would suggest. Interpreting specific nonclassical probabilities as constraints on rational belief is likely to be problematic for analogous reasons.

Now, perhaps one could argue that in the end, such doubts manifest a lack of perfect, ideal rationality—so, at least, the dogmatist must argue. I find this somewhat implausible. To begin with, it may be (as Putnam argued long ago) that the issue of which logic is correct is a broadly empirical one. Whatever one thinks of quantum logic as an putative exemplar of such rationality, I would be surprised if we were never faced with scientific theory choice between total theories embedding incompatible logico-semantic packages. There's certainly a proliferation of such packages available in metaphysics; and many of us are naturalists enough to think that empirical evidence just as much as a priori reflection is holistically relevant to theory choice in metaphysics. It would seem to me a strike against a theory of rational belief if it can't represent rational uncertainty between such physical or metaphysical systems, and the gradual accumulation of evidence for one or the other. But set such considerations aside. The dogmatists need to argue, not just that there is an *absence* of possible evidence in favour of unfavoured logics, but that in some ideal limit there is *positive* evidence for their favoured system, sufficiently strong to require rational certainty. Why think that our total evidence, and superhuman processing power, would convey *total* conviction in the correctness of classical logic, or indeed any other logical system?

This uncertainty challenges all the forms of probabilism we have been discussing. Uncertainty over what the right logic is can lead to attitudes to individual sentences that are condemned by *all* the probability theories based on logics one is open to. Suppose that I divide my credence 50/50 over L-worlds and L^* worlds. And suppose S is true at all L worlds and false at all L^* worlds. Then I should be 50/50 in S. But this is condemned by the 'rational constraints' associated with L and by L^*—L says I should have credence 1; and L^* says I should have credence 0.

So what can we offer an anti-dogmatist? One possibility is to drop the assumption that there is a space of truth-value distributions (classical or otherwise) over which to define probabilities, independently of one's doxastic state. Perhaps the theory of subjective probability should be developed relative to a set of truth-value distributions Z, *that the ideal agent regards as open possibilities*. The arguments above can be used to characterize convex combinations of truth value over the possibilites in Z, and a consequence relation defined via the no drop characterization used before. If the open possibilities include as many varieties of truth-value distributions as has been suggested, then the Z-logic will be weak indeed, and the constraints on rational degrees of belief also weak. But it is a virtue of the framework we've been using that it applies even to this radically minimal setting, and many of the results carry over.[36]

Suppose we decided to theorize about ideally rational belief in this way. Even if the logic and constraints on rational belief are radically minimal, perhaps the majority of a sensible person's credence will be devoted to some $C \subset Z$ which contains—say—only classical truth-value distributions. And if rational degrees of belief have to be convex combinations of truth values, then we do get the non-trivial result that the degrees of belief *conditional on* C have to meet the classical constraints. *Mutatis mutandis* for other interesting regions of the open possibilities Z—the Kleene possibilities K, say. So even though the statable constraints on ideally rational unconditional belief may be rather minimal, it implicitly inherits much richer constraints of rationality, in that it must be such that the updated probabilities $P_C(\cdot)$ be classical probabilities; $P_K(\cdot)$ be Kleene probabilities and so forth.

References

Choquet, G. (1953) Theory of capacities. *Annales e l'nstitute Fourier*. V. pp. 131–295.

de Finetti, Bruno (1974) *Theory of Probability*. Vol. 1. Translated from the Italian by Machi and Smith. New York, NY: Wiley.

Di Nola, A., Georgescu, G. and Lettieri, A. (1999) Conditional states in finite-valued logic. In Dubois, D., Prade, H. and Klement, E. P. (eds.) *Fuzzy Sets, Logics, and Reasoning About Knowledge*. pp. 161–74. Dordrecht: Kluwer.

Dummett, Michael A. E. (1959) Truth. In *Truth and Other Enigmas*. pp. 1–24. London: Duckworth.

[36] Here is one important limitation (thanks to Mark Jago and Mike Caie for pressing me on this). I earlier mentioned and set aside nonclassical settings where truth values are not linearly ordered, or not representable by real numbers. It's not so clear what a convex combination of truth values is to be in that setting; nor whether our results generalize to it. But if these more radical nonclassical probabilities are included in Z, then these issues arise for what is rational to believe relative to Z itself.

Dummett, Michael A. E. (1991) *The Logical Basis of Metaphysics*. The William James lectures; 1976. Cambridge, MA: Harvard University Press.

Field, Hartry (2009) What is the normative role of logic? *Proceedings of the Aristotelian Society*. 83. 3. pp. 251–68.

Field, Hartry H. (2000) Indeterminacy, degree of belief, and excluded middle. *Noûs*. 34. pp. 1–30.

Field, Hartry H. (2003a) No fact of the matter. *Australasian Journal of Philosophy*. 81. pp. 457–80.

Field, Hartry H. (2003b) Semantic paradoxes and the paradoxes of vagueness. In Beall, J. C. (ed.) *Liars and Heaps*. pp. 262–311. Oxford: Oxford University Press.

Field, Hartry H. (2008) *Saving Truth from Paradox*. Oxford: Oxford University Press.

Fine, K. (1975) Vagueness, truth and logic. *Synthese*. 30. pp. 265–300. Reprinted with corrections in Keefe and Smith (eds.) (1997). *Vagueness: A Reader*. pp. 119–50. Cambridge, MA: MIT Press.

Gerla, B. (2000) MV-algebras, multiple bets and subjective states. *International Journal of Approximate Reasoning*. 25. 1. pp. 1–13.

Haack, S. (1978) *Philosophy of Logics*. Cambridge: Cambridge University Press.

Hájek, A. (2008a) Arguments for—or against—probabilism? *British Journal for the Philosophy of Science*. 59.

Hájek, A. (2008b) Dutch book arguments. In Anand, P., Pattanaik, P., and Puppe, C. (eds.) *The Oxford Handbook of Rational and Social Choice*. pp. 173–95. Oxford: Oxford University Press.

Hájek, A. S. (2012) Interpretations of probability. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/archives/sum2012/entries/probability-interpret/ [Accessed 12 Aug 2015].

Hájek, P. (1998) *Metamathematics of Fuzzy Logic*. Dordrecht: Springer.

Harman, G. (1983) Problems with probabilistic semantics. In Orenstein, A., and Stern, R. (eds.) *Developments in Semantics*. pp. 242–35. New York: Haven.

Howson, C. (2008) De Finetti, countable additivity, consistency and coherence. *The British Journal for the Philosophy of Science*. 59. 1. pp. 1–23.

Hughes, R. I. G. (1992) *The Structure and Interpretation of Quantum Mechanics*. Cambridge, MA: Harvard University Press.

Jaffray, J-Y. (1989) Coherent bets under partially resolving uncertainty and belief functions. *Theory and Decision*. 26. pp. 90–105.

Jansana, R. (2011) Propositional consequence relations and algebraic logic. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/consequence-algebraic/ [Accessed 12 Aug 2015].

Jeffrey, R. C. (1983) *The Logic of Decision*. 2nd ed. Chicago and London: University of Chicago Press.

Joyce, J. M. (1998) A non-pragmatic vindication of probabilism. *Philosophy of Science*. 65. pp. 575–603.

Joyce, J. M. (2005) How probabilities reflect evidence. *Philosophical Perspectives*. 19.

Joyce, J. M. (2009) Accuracy and coherence: prospects for an alethic epistemology of partial belief. In Huber, F. and Schmidt-Petri, C. (eds.) *Degrees of Belief*. pp. 263–97. Dordrecht: Springer.

Keefe, R. (2000) *Theories of Vagueness*. Cambridge: Cambridge University Press.

Kraft, C., Pratt, J. and Seidenberg, A. (1959) Intuitive probability on finite sets. *Annals of Mathematical Statistics*. 30. pp. 408–19.

Lewis, D. K. (1980) A subjectivist's guide to objective chance. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability*. Vol. II. pp. 263–93. Berkeley, CA: University of California Press.

Lewis, D. K. (1986) *On the Plurality of Worlds*. Oxford: Blackwell.

MacFarlane, J. G. (2010) Fuzzy epistemicism. In Dietz, Richard, and Moruzzi, Sebastiano (eds.) *Cuts and Clouds*. Oxford: Oxford University Press.

Maudlin, T. (2004) *Truth and Paradox: Solving the Riddles*. Oxford: Oxford University Press.

Milne, P. (2007) Betting on fuzzy and many-valued propositions (long version). In ms.

Milne, P. (2009) Betting on fuzzy and many-valued propositions. In Michal, Peliš (ed.) *The Logica Yearbook 2008*. pp. 145–54. London: College Publications.

Mundici, D. (2006) Bookmaking over infinite-valued events. *International Journal of Approximate Reasoning*. 43. 3. pp. 223–40.

Paris, J. B. (2001) A note on the Dutch Book method. *International Symposium on Imprecise Probabilities and Their Applications*, Ithaca, NY. Maastricht: Shaker.

Priest, G. (2001) *An Introduction to Non-Classical Logic*. Cambridge: Cambridge University Press.

Priest, G. (2006) *In Contradiction: A Study of the Transconsistent*. 2nd ed. New York, NY: Oxford University Press.

Roeper, P. and Leblanc, H. (1999) *Probability Theory and Probability Logic*. Toronto: University of Toronto Press.

Scott, D. (1964) Measurement structures and linear inequalities. *Journal of Mathematical Psychology*. 1. 2. pp. 233–47.

Shafer, G. (1976) *A Mathematical Theory of Evidence*. Princeton, NJ: Princeton University Press.

Smith, Nicholas J. J. (2008) *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.

Smith, Nicholas J. J. (2010) Degree of belief is expected truth value. In Dietz, Richard, and Moruzzi, Sebastiano (eds.) *Cuts and Clouds*. Oxford: Oxford University Press.

Soames, S. (1989) Semantics and semantic competence. *Philosophical Perspectives*. 3. pp. 575–96.

Stalnaker, R. (1984) *Inquiry*. Cambridge, MA: MIT Press.

van Fraassen, B. (1966) Singular terms, truth-value gaps, and free logic. *The Journal of Philosophy*. 63. 17. pp. 481–95.

van Fraassen, B. (1984) Belief and the will. *The Journal of Philosophy*. 81. 5. pp. 235–56.

Weatherson, B. (2003) From classical to constructive probability. *Notre Dame Journal of Formal Logic*. 44. pp. 111–23.

Weatherson, B. (2005) True, Truer, Truest. *Philosophical Studies*. 123. pp. 47–70.

Williams, J. R. G. (2008) Supervaluation and logical revisionism. *The Journal of Philosophy*. 105. 4. pp. 192–212.

Williams, J. R. G. (2012a) Generalized probabilism: Dutch books and accuracy domination. *Journal of Philosophical Logic*. 41. 5. pp. 811–40.

Williams, J. R. G. (2012b) Gradational accuracy and non-classical semantics. *Review of Symbolic Logic*. 5. 4. pp. 513–37.

Williamson, T. (1994) *Vagueness*. London: Routledge.

Williamson, T. (2000) *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wright, C. (2001) On being in a quandary. *Mind*. 110. pp. 45–98.

Zadeh, L. A. (1968) Probability measures of Fuzzy events. *Journal of Mathematical Analysis and Applications*. 23. 2. pp. 421–27.

CHAPTER 13

A LOGIC OF COMPARATIVE SUPPORT: QUALITATIVE CONDITIONAL PROBABILITY RELATIONS REPRESENTABLE BY POPPER FUNCTIONS

JAMES HAWTHORNE

13.1 Introduction

Underlying the usual numerical conception of probability is a more basic qualitative notion, that of *comparative probability*. This comparative notion is formally expressed by weak (partial) order relations among sentences or propositions of the form "A $\succcurlyeq$ B", read "A *is at least as probable as* B". These relations may be employed to represent the *comparative confidence relations* for idealized agents. Interpreted this way, a relation of form "A $\succcurlyeq_\alpha$ B" says that agent α is at least as confident that A is true as that B is true.

Each comparative probability relation $\succcurlyeq$ that obeys certain reasonable constraints (expressed as axioms) can be represented by a corresponding probability function P—i.e. it can be proved that A $\succcurlyeq$ B holds just when P[A] $\geq$ P[B] holds, provided the relation $\succcurlyeq$ is a *complete* (rather than *partial*) order. (See de Finetti 1937/1964; 1974, and Savage 1956.) Thus, one possible answer to the problem of where a Bayesian agent's numerical degrees of belief come from is this: the agent is more confident in some claims than in others, and numerical probabilities merely provide a computationally convenient way of modeling these comparative confidence relations. Furthermore, when the comparative relation $\succcurlyeq$ is only a *partial order* it will instead be representable by a set of precise probability functions, each extending $\succcurlyeq$ to a complete order—where for each representing probability function P: (1) whenever A $\succcurlyeq$ B and *not* B $\succcurlyeq$ A, P[A] >P[B], and (2) whenever A $\succcurlyeq$ B and B $\succcurlyeq$ A, P[A] = P[B].

This comparative notion of probability cannot capture some important probabilistic concepts, such as the concept of probabilistic dependence and independence. This deficiency may be remedied by extending the comparative concept to a notion of *comparative conditional probability*. A *comparative conditional probability relation* is a weak (partial) order among pairs of sentences or propositions of form "A|B $\succcurlyeq_\alpha$ C|D", read "A, given B, *is at least as probable as* C, given D".[1]

These relations may be employed to represent idealized agents' *comparative conditional confidence relations*. Interpreted this way, a relationship of form "A|C $\succcurlyeq_\alpha$ C|D" says that agent α is at least as confident that A (is true), given that B, as she is that C, given that D. However, an alternative (arguably distinct) conception employs these relations to represent *comparative argument strengths*. Interpreted this way, a relationship of form "A|B $\succcurlyeq_\alpha$ C|D" says that under an interpretation α of the inferential import (or inferential meanings) of statements of the language, conclusion A is supported by premise B at least as strongly as conclusion C is supported by premise D. J. M. Keynes suggested this kind of reading of conditional probabilities in his *Treatise on Probability* (1921). B. O. Koopman (1940) axiomatized this Keynesian conception in terms of a logic of *comparative conditional probability* of the kind developed below.

In this chapter I will present axioms for *comparative conditional probability relations* that are more general than usual. Each of these relations is a weak partial order on pairs of sentences—i.e. each relation will be transitive and reflexive, but need not be a complete order relation. The axioms presented here are *probabilistically sound* for the broad class of conditional probability functions known as *Popper functions* (which will be axiomatized in Section 13.2).[2] That is, for each Popper function P, the corresponding *comparative conditional probability relation* $\succcurlyeq$ (defined by "A|B $\succcurlyeq$ C|D" whenever P[A|B] $\geq$ P[C | D]) will satisfy the axioms below for *comparative conditional probability relations*. Furthermore, these axioms are *probabilistically complete*: a representation theorem shows that for each relation $\succcurlyeq$ that satisfies these axioms, there is a corresponding Popper function P such that, for all sentences A, B, C, D: (i) whenever the relationship A|B $\succ$ C|D holds (i.e. whenever A|B $\succcurlyeq$ C|D but *not* C|D $\succcurlyeq$ A|B), the corresponding probabilistic relationship P[A|B] >P[C|D] holds; (ii) whenever A|B $\approx$ B|D holds (i.e. whenever A|B $\succcurlyeq$ C|D and C|D $\succcurlyeq$ A|B), the corresponding probabilistic relationship P[A|B] = P[C|D] holds.[3]

[1] Two examples: [the coin comes up *heads* in this case|the coin is fair and flipped properly in this case] *is at least as probable as* [the die lands *six* on this toss|the die is fair and tossed properly on this toss]; [it will rain here later today|the barometer fell rapidly earlier today] *is at least as probable as* [the Democrats win a senate seat in Arizona next election|there is no major change in party politics in the US before the next election].

[2] Koopman's (1940) axiomatization leaves relationships A|B $\succcurlyeq$ C|D undefined whenever B (or D) has "0-probability"—i.e. whenever (E · ¬E)|(E∨¬E) $\succcurlyeq$ B|(E∨¬E). However, Popper functions permit P[A|B] to have well-defined values between 0 and 1 even in cases where P[B|(E∨¬E)] = 0. The axioms for the comparative relations provided below follow suit by permitting "A|B $\succcurlyeq$ C|D" to remain defined for all pairs of sentence pairs.

[3] More generally, a comparative relation $\succcurlyeq$ may be representable by a set of distinct Popper functions that disagree on numerical values, but agree on the orderings among conditional probabilities. This provides an entry into theories of imprecise and indeterminate probabilities (Koopman 1940). A detailed account is provided in the article by Fabio Cozman in Chapter 14 in this volume.

The axiomatic system I'll present is purely formal. So the *comparative conditional probability relations* the axioms characterize may be interpreted in terms of any of the usual probabilistic concepts. For example, one may interpret these relations in terms of some notion of *comparative conditional chance*. On this sort of interpretation a relationship of form "A|B $\succcurlyeq_\alpha$ C|D" may be read as "for systems in state α, the chance of outcome A among those systems (or states of affairs) with attribute B *is at least as great as* the chance of outcome C among those with attribute D". On this reading the representation theorems will show that the usual numerical conditional chance functions provide a convenient way to represent a purely qualitative-comparative conception of conditional chance relations among states of affairs.

Although the abstractness of the formalism provides generality, the axioms for comparative conditional probability relations will be easier to motivate if we give the comparative relations $\succcurlyeq$ some uniform interpretive reading throughout. So, henceforth I'll read each such comparative relation as expressing comparisons among arguments with respect to support-strength. Each relationship "A|B $\succcurlyeq$ C|D" will be read as "conclusion A is supported by the conjunction of premises B *at least as strongly as* conclusion C is supported by the conjunction of premises D". Thus, henceforth we will be investigating *comparative conditional probability* as a logic of *comparative argument strength*, a qualitative logic which may provide a foundation for the Bayesian logic of evidential support. Readers interested in other conceptions of probability are invited to see how well those conceptions fit the axioms on offer here.

13.2 Popper Functions

Popper functions are a generalization of the usual classical notion of conditional probability. All classical conditional probability functions are (in effect) very restricted kinds of Popper functions—i.e. they satisfy the axioms for Popper functions, provided that in cases where classical conditional probabilities are left undefined, we define them as equal to 1.[4] However, among the Popper functions are conditional probability functions that make important use of conditionalization on statements that have probability 0. I'll say more about this below.

Various axiomatizations of the Popper functions are available. Karl Popper's original motivation was to develop a probabilistic logic that does not presuppose (and does not draw on) classical deductive logic, and to then show that classical deductive logical entailment arises as a special case of a purely probabilistic notion of entailment.[5] I'll bypass this aspect

[4] A *classical probability function* on language *L* (for sentential or predicate logic) is any function p from sentences to real numbers between 0 and 1 that satisfies the following axioms: (1) if|= A, then p[A] = 1; (2) if|= ¬(A · B), then p[(A∨B)] = p[A] + p[B]; (3) (definition) when p[B] >0, p[A|B] = p[(A · B)] / p[B]. When p[B] = 0, p[A|B] is undefined (or, may be defined to equal 1). These axioms suffice. It follows from them that when B | = A, p[B] ≤ p[A] (given B|= A,|= ¬(B · ¬A), so 1 ≥ p[(B∨¬A)] = p[B] + p[¬A], so 1 − p[¬A] ≥ p[B]; and since | = ¬(A · ¬A) and|= (A∨¬A), 1 = p[(A∨¬A)] = p[A] + p[¬A]; so p[A] = 1 − p[¬A] ≥ p[B]). From this it follows that logically equivalent sentences have the same probability.

[5] See the appendix to Popper (1959). Hartry Field (1977) shows how to extend Popper's project to probability functions for predicate logic. That is, Field shows how to construct a probabilistic semantics

of Popper's project here, and build the logic of the Popper functions atop classical deductive logic.

Popper functions turn out to have another important feature. They provide a significant way to generalize the classical notion of conditional probability. I'll say more about this in a while. First, here is a fairly sparse way to axiomatize the Popper functions. These particular axioms are informative because, weak as they are, they provide close analogues of the axioms for *comparative conditional probability relations* introduced later. The following axioms only suppose that numerical values are real numbers—the restriction to values between 0 and 1 must be proved.

> *Sparse Axioms for Popper Functions*: Let *L* be a language having either the syntax of sentential logic, or alternatively, the syntax of predicate logic (including identity and functions).[6] Let '|=' represent the usual logical entailment relation for the logic (either sentential logic or predicate logic). Each Popper function is a function P from pairs of sentences of *L* to the real numbers such that for all sentences A, B, and C:
>
> 1. for some sentences E, F, G, H, $P[E|F] \neq P[G|H]$
> 2. $P[A|A] \geq P[B|C]$
> 3. if $B| = A$, then $P[A|C] \geq P[B|C]$
> 4. if $C| = B$ and $B| = C$, then $P[A|B] \geq P[A|C]$
> 5. $P[A|B] + P[\neg A|B] = P[B|B]$ or else $P[D|B] = P[B|B]$ for all D
> 6. $P[(A \cdot B)|C] = P[A|(B \cdot C)] \times P[B|C]$

This axiomatization is so weak that the usual probabilistic formulae are difficult to derive. It is useful for our purposes because of its close connection with the axioms for *comparative conditional probability relations* provided later. Here is an alternative, more usual axiomatization of the Popper functions.

> *Robust Axioms for Popper Functions*: Let *L* be a language having the syntax of either sentential logic or predicate logic (including identity and functions), where '|=' represents the usual logical entailment relation. Each Popper function is a function P from pairs of sentences of *L* to the real numbers such that for all sentences A, B, and C:
>
> (1) if $| = \neg A$ and $| = B$ (i.e. A is a contradiction and B is a tautology), then $P[A|B] = 0$
> (2) $1 \geq P[A|B] \geq 0$
> (3) if $B| = A$, then $P[A|B] = 1$
> (4) if $C| = B$ and $B| = C$, then $P[A|B] = P[A|C]$
> (5) if $C| = \neg(A \cdot B)$, then either $P[(A \vee B)|C] = P[A|C] + P[B|C]$ or $P[D|C] = 1$ for all D
> (6) $P[(A \cdot B)|C] = P[A|(B \cdot C)] \times P[B|C]$.

Clearly, the *sparse axioms*, 1–6, are derivable from the *robust axioms*, (1)–(6). The derivation of the robust axioms from the sparse axioms requires some effort (see the Appendix).

for predicate logic that takes the notion of probability assignments (rather than truth-value assignments) as basic. He proves that this semantics gives rise to a notion of logical entailment that is coextensive with the classical notion.

[6] Probabilistic logics are often restricted to a language for sentential logic, but everything here carries over to full predicate logic with identity and functions.

To understand the relationship between Popper functions and classical conditional probability functions, think of it like this. Given an *unconditional* classical probability function p, *conditional probability* is usually defined as follows: whenever $p[B] > 0$, $p[A|B] = p[(A \cdot B)]/P[B]$; and when $p[B] = 0$, $p[A|B]$ is left undefined. Let's make a minor modification to this usual approach, and require instead that classical conditional probability functions make $p[A|B] = 1$ by default whenever $p[B] = 0$. Thus, on this approach conditional probabilities are always defined. Specified in this way, each classical conditional probability function is a simple kind of Popper function (i.e. it satisfies the axioms for Popper functions).

More generally, a Popper function may consist of a ranked hierarchy of classical probability functions, where conditionalization on a probability 0 sentence induces a transition from one classical probability function to another classical function at a lower rank. The idea is that probability 0 need not mean "absolute impossibility". Rather, it means something like, "not a viable possibility unless (and until) the more plausible alternatives are refuted".

Here is how that works in more detail. For a given Popper function P, if we hold the condition statement B fixed, then the function $P[|B]$ behaves precisely like a classical probability function—it always satisfies the classical axioms. However, when a statement C has 0 probability on B, $P[C|B] = 0$, the probability function $P[\ |\ (C \cdot B)]$ gotten by now holding the conjunction $(C \cdot B)$ fixed may remain well-defined, and may behave like an entirely different classical probability function. In general, a Popper function consists of a ranked hierarchy of classical probability functions, where the transition from a classical probability function at one level in the hierarchy (the statement B level) to a new classical probability function at a lower level (the statement $(C \cdot B)$ level) is induced by conditionalization on a statement $(C \cdot B)$ that has probability 0 at that higher (statement B) level. Finally, at the bottom level, below all other ranks associated with P, is the level of logical contradictions. This level may also include sentences that "behave like logical contradictions"—i.e. sentences E such that every sentence has probability 1 when conditionalized on E: $P[A|E] = 1$ for all A.

The fact that a Popper function may consist of this kind of ranked hierarchy of classical functions is not an additional assumption or stipulation. Rather, it follows from the above axioms (from 1–6, and also from (1)–(6)) without supplementation. (See Hawthorne 2014, for a detailed account of the ranked structure of Popper functions, including proofs of these claims.)

Here is an illustration of a case where this kind of generalization of classical probability proves useful. Suppose that the probability that a randomly selected point will lie within the upper 3/4 of a specific spatial region, described by "A", given that it lies somewhere within that whole three-dimensional region, described by "B", is $P[A|B] = 3/4$. The probability that this same randomly selected point will lie precisely on a specific plane described by "C" where it intersects the B-region should presumably be 0, so we have $0 = P[C|B] = P[(A \cdot C)|B]$. However, given that this random point does indeed lie within the C-plane within the B-region, the probability that it lies within the part of that region described by "A" (which, say, contains half of the plane described by '$C \cdot B$') may again be perfectly well-defined: $P[A\ |(C \cdot B)] = 1/2$. Furthermore, the probability that this random point will lie on the part of a line segment described by "D" within the C-plane should also presumably be 0, so we again have a situation where $0 = P[D|(C \cdot B)] = P[(A \cdot D)|(C \cdot B)]$. However, given

that this random point does indeed lie within the part of the D-segment within the part of the C-plane within the B-region, the probability that it lies in the A-region (which, say, contains two-thirds of the D-line-segment within (D · (C · B))) may again be well-defined: $P[A|(D \cdot (C \cdot B))] = 2/3$. So, the general idea is that a specific Popper function may consist of a ranked hierarchy of classical probability functions, where conditionalizations on specific probability 0 statements at one level of the hierarchy can induce a transition to another perfectly good classical probability function defined at a lower level of the hierarchy.[7]

Bayesian confirmation theory employs conditional probability functions to represent the support of evidence for hypotheses, and Popper functions may serve in this role. However, the Bayesian approach to confirmation owes us an account of where the proposed numerical *degrees of support* come from, and what the probabilistic numbers mean or represent. Subjectivist Bayesians attempt to provide this account in terms of betting functions and Dutch book theorems—they maintain that confirmation functions are belief-strength functions, and that their numerical values represent ideally rational betting quotients, which must satisfy the usual probabilistic rules in order to avoid the endorsement of betting packages that would result in sure losses. However, for a logical account of confirmation functions (e.g. of the kind endorsed by Keynes), wherein confirmation functions represent argument strengths, another kind of answer to the "where do the numbers come from, and what do they mean?" question may be offered—an answer via a *representation theorem*. On this approach the idea is that confirmation theory derives from a deeper underlying *qualitative logic of comparative argument strength*. We now proceed to specify the rules that govern this deeper logic. We'll see that Popper functions merely provide a convenient way to calculate the comparative support relationships captured by this qualitative logic of comparative argument strength; the probabilities add nothing that the qualitative logic cannot already capture on its own.

13.3 Towards the Logic of *Comparative Argument Strength*: the *Proto-Support Relations*

A *comparative support relation* $\succcurlyeq$ is a relation among pairs of sentence pairs. It should satisfy axioms that provide plausible restrictions on a reasonable conception of the notion of *comparative argument strength*. We will get to the axioms in a moment.

Associated with each relation $\succcurlyeq$ are several related relations, defined in terms of it. Here is a list of these, their formal definitions, and an appropriate informal reading for each.

A comparative support relation $\succcurlyeq$ is a relation of form $A|B \succcurlyeq C|D$,
read "A *is supported by* B *at least as strongly as* C *is supported by* D".

[7] For a comparison of the Popper functions to other accounts of conditional probability functions, see Chapter 9 by Kenny Easwaran in this volume.

Define four associated relations as follows:

(1) A|B $\succ$ C|D abbreviates "A|B $\succcurlyeq$ C | D and *not* C|D $\succcurlyeq$ A|B",
read "A *is supported by* B *more strongly than* C *is supported by* D";
(2) A|B $\approx$ C|D abbreviates "A|B $\succcurlyeq$ C|D and C|D $\succcurlyeq$ A|B",
read "A *is supported by* B *to the same extent that* C *is supported by* D";
(3) A|B $\asymp$ C|D abbreviates "*not* A|B $\succcurlyeq$ C|D and *not* C|D $\succcurlyeq$ A|B",
read "*the support for* A *by* B *is indistinctly comparable to that of* C *by* D";
(4) B $\Rightarrow$ A abbreviates "A|B $\succcurlyeq$ C|C";
read "B *supportively entails* A".

The axioms I'll provide for $\succcurlyeq$ turn out to entail that for each such relation, the corresponding *supportive entailment relations* $\Rightarrow$ satisfies the rules for a well-known kind of non-monotonic conditional called a *rational consequence relation*. Indeed, the *rational consequence relations* turn out to be identical to the *supportive entailment relations*.[8]

With these definitions in place we are ready to specify axioms for the *comparative support relations*. Axioms 1-6 closely parallel the corresponding axioms for Popper functions. The axioms will only ensure that these comparative relations are partial orders on comparative argument strength: they are transitive and reflexive, but need not be complete orders—i.e. some argument pairs may fail to be distinctly comparable in strength.

Let *L* be a language having the syntax of sentential logic or predicate logic (including identity and functions). Each *proto-support relation* $\succcurlyeq$ is a binary relation between pairs of sentences that satisfies the following axioms:

0. If A|B $\succcurlyeq$ C|D and C|D $\succcurlyeq$ E|F, then A|B$\succcurlyeq$ E|F (transitivity)
1. for some E, F, G, H, E|F $\succ$ G|H (non-triviality)

[Not all arguments are equally strong; at least one is stronger than at least one other.]

2. A|A $\succcurlyeq$ B|C (maximality)

[Self-support is maximal support – at least as strong as any argument.]

3. If B|= A, then A|C $\succcurlyeq$ B|C (classical consequent entailment)

[Whenever B logically entails A, *the support for* A *by* C *is at least as strong as the support for* B *by* C. The reflexivity of $\succcurlyeq$ follows, since A|= A. Together with other axioms it yields: (i) If (B · C)|= A, then A|C $\succcurlyeq$ B|C; (ii) If B|= A, then A|B $\succcurlyeq$ C|D.]

4. If B|= C and C|= B, then A|B $\succcurlyeq$ A|C (classical antecedent equivalence)

[8] The ranked structure of the *Popper functions* (structured as a hierarchy of classical probability functions) is just the ranked structure of the rational consequence relations. The comparative support relations turn out to share this ranked structure, captured by their associated supportive entailment relations. See Hawthorne (2014) for a detailed account of the rational consequence relations and their ranked structures.

[Logically equivalent statements support all statements equally well.]

5. If $A|B \succcurlyeq C|D$, then $\neg C|D \succcurlyeq \neg A|B$ or else $B \Rightarrow D$ for all D (negation-symmetry)

[Whenever A *is supported by* B *at least as strongly as* C *is supported by* D, *the falsity of* C *is supported by* D *at least as strongly as the falsity of* A *is supported by* B; the only exception is in cases where premise B behaves like a contradiction, maximally supporting every statement D. This captures the essence of the additivity axiom for conditional probability, axiom 5 for the Popper functions.]

6.1 If $H_1|(A_1 \cdot E_1) \succcurlyeq H_2|(A_2 \cdot E_2)$ and $A_1|E_1 \succcurlyeq A_2|E_2$, then $(H_1 \cdot A_1)|E_1 \succcurlyeq (H_2 \cdot A_2)|E_2$

6.2 If $H_1|(A_1 \cdot E_1) \succ H_2 \mid (A_2 \cdot E_2)$ and $A_1|E_1 \succcurlyeq A_2|E_2$, then either $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$

6.3 If $H_1|(A_1 \cdot E_1) \succcurlyeq H_2 \mid (A_2 \cdot E_2)$ and $A_1|E_1 \succ A_2|E_2$, then either $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$

6.4 If $H_1|(A_1 \cdot E_1) \succcurlyeq A_2|E_2$ and $A_1|E_1 \succcurlyeq H_2|(A_2 \cdot E_2)$, then $(H_1 \cdot A_1)|E_1 \succcurlyeq (H_2 \cdot A_2)|E_2$

6.5 If $H_1|(A_1 \cdot E_1) \succ A_2 \mid E_2$ and $A_1|E_1 \succcurlyeq H_2|(A_2 \cdot E_2)$, then either $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$

6.6 If $H_1|(A_1 \cdot E_1) \succcurlyeq A_2|E_2$ and $A_1|E_1 \succ H_2|(A_2 \cdot E_2)$, then either $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$

[These *composition* axioms, 6.1–6.6, together with the following four axioms (6.7–6.10) capture the essential content of probabilistic conditionalization, expressed by axiom 6 for the Popper functions. Think of H_1 and H_2 as hypotheses, A_1 and A_2 as auxiliary hypotheses, and E_1 and E_2 as evidence statements. For $k = 1,2$, $P[H_k \cdot A_k|E_k] = P[H_k \mid A_k \cdot E_1] \times P[A_k|E_k]$, so when $P[H_1|A_1 \cdot E_1] \geq P[H_2 \mid A_2 \cdot E]$ and $P[A_1|E_1] \geq P[A_2 \mid E_2]$, we must have $P[H_1 \cdot A_1|E_1] \geq P[H_2 \cdot A_2|E_2]$, with $P[H_1 \cdot A_1|E_1] > P[H_2 \cdot A_2 \mid E_2]$ when either $P[H_1|A_1 \cdot E_1] > P[H_2|A_2 \cdot E]$ or $P[A_1|E_1] > P[A_2|E_2]$. The "or clause" in 6.2 provides that when $\neg A_1$ has probability 1 on E_1 (so A_1 has probability 0 on E_1), "$(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2) \mid E_2$" need not hold—that is, when $0 = P[A_1|E_1] \geq P[A_2|E_2]$, we must have $0 = P[(H_1 \cdot A_1)|E_1] = P[(H_2 \cdot A_2) \mid E_2]$. Clauses 6.3, 6.5, and 6.6 may be explained similarly.]

6.7 If $H_1|(A_1 \cdot E_1) \asymp H_2|(A_2 \cdot E_2)$ and $A_1|E_1 \succcurlyeq A_2|E_2$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$

6.8 If $H_1|(A_1 \cdot E_1) \succcurlyeq H_2|(A_2 \cdot E_2)$ and $A_1|E_1 \asymp A_2|E_2$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$

6.9 If $H_1|(A_1 \cdot E_1) \asymp A_2|E_2$ and $A_1|E_1 \succcurlyeq H_2|(A_2 \cdot E_2)$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$

6.10 If $H_1|(A_1 \cdot E_1) \succcurlyeq A_2|E_2$ and $A_1|E_1 \asymp H_2|(A_2 \cdot E_2)$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$

[In the presence of axiom 6.1 the conjunction of axioms 6.2 and 6.7 is equivalent to the following rule, called "decomposition" (i.e. the latter two axioms could be replaced by it):

If $(H_2 \cdot A_2)|E_1 \succcurlyeq (H_1 \cdot A_1)|E_2$ and $A_1|E_1 \succcurlyeq A_2|E_2$ and $E_1 \not\Rightarrow \neg A_1$, then $H_2|(A_2 \cdot E_2) \succcurlyeq H_1|(A_1 \cdot E_1)$.

Similarly, in the presence of axiom 6.1, the conjunction of 6.3 and 6.8 is equivalent to the following decomposition rule:
If $(H_2 \cdot A_2)|E_1 \succcurlyeq (H_1 \cdot A_1)|E_2$ and $H_1|(A_1 \cdot E_1) \succcurlyeq H_2|(A_2 \cdot E_2)$ and $E_1 \not\Rightarrow \neg H_1$, then $A_2|E_2 \succcurlyeq A_1|E_1$.

In the presence of axiom 6.4, the pairs of axioms {6.5, 6.9} and {6.6, 6.10}are equivalent to corresponding decomposition rules.][9]

7. If $A|(B \cdot C) \succcurlyeq E|F$ and $A|(B \cdot \neg C) \succcurlyeq E|F$, then $A|B \succcurlyeq E|F$ (alternate presumption)*

[Probabilistically this axiom follows from additivity together with conditionalization—i.e. since $P[A|B] = P[A|B \cdot C] \times P[C|B] + P[A|B \cdot \neg C] \times (1 - P[B|C])$, if both $P[A|B \cdot C] \geq r$ and $P[A|B \cdot \neg C] \geq r$, then $P[A|B] \geq r$. Axiom 7 is a qualitative version of this result. It can be proved from the other axioms if the relation $\succcurlyeq$ is assumed to be a *complete order* rather than merely a partial order relation—i.e. if $\succcurlyeq$ takes all argument pairs to be distinctly comparable in strength.]

All relations that satisfy these axioms are weak partial orders—i.e. they are transitive and reflexive. Transitivity is guaranteed by axiom 0; reflexivity, $A|C \succcurlyeq A|C$, follows from axiom 3. I call the relations that satisfy these axioms *proto-support relations* because the axioms still need a bit of strengthening to rule out some relations $\succcurlyeq$ that fail to behave properly. I'll say more about that later.

The asterisk on the name of axiom 7 is to indicate that it follows from the other axioms whenever the relation $\succcurlyeq$ is also *complete*—i.e. whenever, for all pairs of sentence pairs, the

[9] It may seem odd that the 6.7–6.10 axioms contain $\succcurlyeq$ in their antecedents and $\succ$ in their consequents. Here's what is going on. 6.7 is equivalent to the following two rules:

6.7.1 If $H_1|(A_1 \cdot E_1) \asymp H_2 \mid (A_2 \cdot E_2)$ and $A_1|E_1 \succ A_2 \mid E_2$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$
6.7.2 If $H_1 \mid (A_1 \cdot E_1) \approx H_2|(A_2 \cdot E_2)$ and $A_1|E_1 \asymp A_2|E_2$, then $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$

Similarly, 6.8 is equivalent to the following two rules:

6.8.1 If $H_1|(A_1 \cdot E_1) \succ H_2|(A_2 \cdot E_2)$ and $A_1|E_1 \asymp A_2|E_2$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$
6.8.2 If $H_1|(A_1 \cdot E_1) \approx H_2 \mid (A_2 \cdot E_2)$ and $A_1|E_1 \asymp A_2 \mid E_2$, then $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$

The following three rules are equivalent to axioms 6.9–6.10.

6.9.1 If $H_1 \mid (A_1 \cdot E_1) \asymp A_2|E_2$ and $A_1 \mid E_1 \succ H_2|(A_2 \cdot E_2)$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$
6.10.1 If $H_1|(A_1 \cdot E_1) \asymp A_2 \mid E_2$ and $A_1|E_1 \approx H_2|(A_2 \cdot E_2)$, then $(H_1 \cdot A_1)|E_1 \succ (H_2 \cdot A_2)|E_2$ or $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $(A_1 \cdot E_1) \Rightarrow \neg H_1$
6.10.2 If $H_1|(A_1 \cdot E_1) \asymp A_2|E_2$ and $A_1|E_1 \approx H_2|(A_2 \cdot E_2)$, then $(H_1 \cdot A_1)|E_1 \asymp (H_2 \cdot A_2)|E_2$ or $E_1 \Rightarrow \neg A_1$.

following *complete comparability rule* also holds (or is added as an additional axiom) for relation $\succcurlyeq$:

> either $A|B \succcurlyeq C|D$ or $C|D \succcurlyeq A|B$ (complete comparability).

Adding the rule for *complete comparability* would require that all argument pairs are distinctly comparable in strength: for all A, B, C, D, $A|B \not\asymp C|D$. Any relation $\succcurlyeq$ that satisfies the above axioms together with *complete comparability* is a weak order relation rather than merely a weak partial order.[10] Also notice that in the presence of *complete comparability* axioms 6.7–6.10 are superfluous; their antecedent conditions involving '$\asymp$' are vacuously satisfied.

The above axioms for *proto-support relations* are probabilistically sound in the following sense:

> For each Popper function P, define the *corresponding comparative relation* to be the relation $\succcurlyeq$ such that, for all sentences A, B, C, D, $A|B \succcurlyeq C|D$ holds if and only if $P[A|B] \geq P[C|D]$.
>
> Then for each Popper function, the *corresponding comparative relation* can be shown to be a *proto-support relation*—it satisfies the above axioms.

The axioms for the *proto-support relations* are not *probabilistically complete*. Some *proto-support relations* are not probability-like enough to be representable by a Popper function. Below we add additional constraints (additional axioms) that suffice to characterize the "full" *comparative support relations*. These relations will turn out to be *probabilistically complete* in the sense that each such *comparative support relation* $\succcurlyeq$ is representable by a Popper function P.

The *proto-support relations* are sufficiently strong to provide some comparative forms of Bayes' theorem. Here is one example.

Bayes' Theorem 1: Suppose $B \not\Rightarrow \neg H_1$.

If $E|(H_1 \cdot B) \succ E|(H_2 \cdot B)$ and $H_1|B \succcurlyeq H_2|B$, then $H_1|(E \cdot B) \succ H_2|(E \cdot B)$.

Think of H_1 and H_2 as hypotheses, B as common background knowledge and auxiliary hypotheses, and E as the evidence. This is an analogue of the following version of Bayes' theorem:

Suppose $P[H_1|B] > 0$. Then

$P[H_2|E \cdot B] / P[H_1|E \cdot B] = (P[E|H_2 \cdot B] / P[E|H_1 \cdot B]) \times (P[H_2|B] / P[H_1|B])$,

so, if $P[E|H_1 \cdot B] > P[E|H_2 \cdot B]$ and $P[H_1|B] \geq P[H_2|B]$, then $P[H_1|(E \cdot B)] > P[H_2|(E \cdot B)]$.

Here is a second form of Bayes' theorem satisfied by *proto-support relations*.

Bayes' Theorem 2: Suppose $B \not\Rightarrow \neg H_1$, $B \not\Rightarrow \neg H_2$, and $B \Rightarrow \neg(H_1 \cdot H_2)$.

If $E|(H_1 \cdot B) \succ E|(H_2 \cdot B)$, then

[10] Koopman also provides the following rule as an axiom:

For any integer $n > 1$, if $A_1, ..., A_n$, and $B_1, ..., B_n$ are collections of sentences such that

$C \not\Rightarrow \neg C$, $C \Rightarrow (A_1 \vee ... \vee A_n)$, $C \Rightarrow \neg(A_i \cdot A_j)$, $A_n|C \succcurlyeq \ldots \succcurlyeq A_2|C \succcurlyeq A_1|C$, and

$D \not\Rightarrow \neg D$, $D \Rightarrow (B_1 \vee ... \vee B_n)$, $D \Rightarrow \neg(B_i \cdot B_j)$, $B_n|D \succcurlyeq \ldots \succcurlyeq B_2|D \succcurlyeq B_1|D$,

then $A_n|C \succcurlyeq B_1|D$ (subdivision)*

This rule may not seem as intuitively compelling as the others, so I forego it here. Later we will want to require that comparative support relations be *extendable* to *complete comparability* relations. *Subdivision* is derivable from axioms 0-6.6 in the presence of *complete comparability*.

$H_1|(E \cdot B \cdot (H_1 \vee H_2)) \succ H_1|(B \cdot (H_1 \vee H_2))$ and
$H_2 \mid (B \cdot (H_1 \vee H_2)) \succ H_2 \mid (E \cdot B \cdot (H_1 \vee H_2))$.

A straightforward probabilistic analogue goes like this:

Suppose $P[H_1|B] > 0$, $P[H_2|B] > 0$, and $P[H_1 \cdot H_2|B] = 0$. If $P[E|H_1 \cdot B] > P[E|H_2 \cdot B]$, then $P[H_1|E \cdot B \cdot (H_1 \vee H_2)] > P[H_1|B \cdot (H_1 \vee H_2)]$ and $P[H_2|E \cdot B \cdot (H_1 \vee H_2)] < P[H_2|B \cdot (H_1 \vee H_2)]$. This is a comparative expression of the relationship, $P[H_2 \mid E \cdot B] / P[H_1|E \cdot B] < P[H_2|B] / P[H_1|B]$, since $P[H_2|E \cdot B \cdot (H_1 \vee H_2)] / P[H_1|E \cdot B \cdot (H_1 \vee H_2)] = P[H_2|E \cdot B] / P[H_1 |E \cdot B]$ and $P[H_2|B \cdot (H_1 \vee H_2)] / P[H_1|B \cdot (H_1 \vee H_2)] = P[H_2|B] / P[H_1|B]$.

13.4 The *Comparative Support Relations* and their Probabilistic Representations

Consider the following additional rules.

8. $A|B \succcurlyeq C|D$ or $C|D \succcurlyeq A|B$ (complete comparability)
9. For each integer $m \geq 2$ there is an integer $n \geq m$ such that for n sentences $S_1, ..., S_n$ and some sentence G:

(i) $G \not\Rightarrow \neg S_1$, and for all distinct i, j,
(ii) $G \Rightarrow \neg(S_i \cdot S_j)$ and
(iii) $S_i|G \text{ “ } S_j|G$. (existence of arbitrarily large equal-partitions)

An *equal-partition* (given G) is a collection of pairwise mutually exclusive sentences (the sentences S_k) that are “equally likely” (given G). When sentences G and collection of n sentences, $S_1, ..., S_n$, satisfy rule 9 for relation $\succcurlyeq$, it can be shown that: (i) $(G \cdot (S_1 \vee ... \vee S_n)) \not\Rightarrow \neg S_1$, for distinct i, $j \leq n$, (ii) $(G \cdot (S_1 \vee ... \vee S_n)) \Rightarrow \neg(S_i \cdot S_j)$, (iii) $S_i|(G \cdot (S_1 \vee ... \vee S_n)) \approx S_j|(G \cdot (S_1 \vee ... \vee S_n))$, and (iv) $(G \cdot (S_1 \vee ... \vee S_n)) \Rightarrow (S_1 \vee ... \vee S_n)$. Thus, rule 9 guarantees that for arbitrarily large n, $\succcurlyeq$ has an exclusive and exhaustive equal-partition based on $(G \cdot (S_1 \vee ... \vee S_n))$. These partitions can be used to provide approximate probability values for the strengths of arguments:
when $(S_1 \vee ... \vee S_{k+1})|(G \cdot (S_1 \vee ... \vee S_n)) \succcurlyeq A|B \succcurlyeq (S_1 \vee ... \vee S_k) \mid (G \cdot (S_1 \vee ... \vee S_n))$ we effectively get a probability-like approximation for the strength of A|B, $(k+1)/n \geq P[A|B] \geq k/n$.

For arbitrarily large partitions (for arbitrarily large values of n) these partitions provide arbitrarily close probability-like bounds on the strength of each argument.

Rule 9 is not as strong as needed in most cases. Here is a stronger alternative:

9^+. If $A|B \succ C|D$, then for some $n \geq 2$ there are sentences $S_1, ..., S_n$ and a sentence F such that:
$F \not\Rightarrow \neg S_1$, and for distinct i, j, $F \Rightarrow \neg(S_i \cdot S_j)$ and $S_i|F \text{ “ } S_j|F$, and $F \Rightarrow (S_1 \vee ... \vee S_n)$, and for some m of them, $A|B \succ (S_1 \vee \ldots \vee S_m)|F \succ C|D$. (Archimedean equal-partitions)

Rule 9^+ implies 9, but adds to it a kind of "Archimedean condition": whenever $A|B \succ C|D$, there must be an equal-partition that, for sufficiently large n, squeezes a "strength comparison" between A|B and C|D. This forces these two arguments to exhibit distinct probabilistic values:

$$P[A|B] > m/n > P[C|D].$$

A comparative support relation that satisfies rules 8 and 9 but that fails to satisfy 9^+ must permit some argument pairs for which $A|B \succ C|D$, but where A|B and C|D lie infinitesimally close together in comparative strength—i.e. no segment of any equal-partition argument can fit between them. There are interesting cases where such non-Archimedean support relations are useful. So I'll treat the full range of relations that satisfy rule 9, as well as the better-behaved subclass of Archimedean relations, which satisfy the more restrictive rule 9^+.

Here is an intuitive example of the kind of partition required by rules 9 and 9^+. Let statement F (a.k.a. statement $G \cdot (S_1 \vee ... \vee S_n)$) describe a fair lottery consisting of exactly n tickets. Each of the sentences S_i says "ticket i will win". F *says* via *supportive entailment* (or via *logical entailment*, which is stronger):

(1) "at least one ticket will win": so $F \Rightarrow (S_1 \vee \ldots \vee S_n)$;
(2) "no two tickets will win": so $F \Rightarrow \neg(S_i \cdot S_j)$, for each distinct pair of claims S_i and S_j;
(3) "each ticket has the same chance of winning" (and where the argument from F provides exactly the same support for the claim "ticket i will win" as for the claim "ticket j will win"): so $S_i|F \approx S_j|F$ for each distinct pair of claims S_i and S_j.
(4) furthermore, we suppose that F does not *supportively entail* "ticket 1 won't win": that is, $F \not\Rightarrow \neg S_1$ (formally this clause is equivalent to $F|F \succ \neg S_1|F$; it eliminates the possibility that F behaves like a contradiction, supportively entailing every statement).

We could require all comparative support relations to satisfy rule 9. This would not be too implausible—it would require merely that the language of each relation $\succcurlyeq$ have the ability to describe such lotteries for arbitrarily large finite numbers of tickets. Presumably our own natural language can do that. So this would be a fairly innocuous requirement.[11] Nevertheless, we won't require that comparative support relations employ languages this rich. Rather, it will suffice for our purposes to suppose that each support relation is (in principle) *extendable* to a relation that includes such lottery descriptions. I'll say more about *extendability* in a bit. Before doing so, let's consider rule 8 more closely.

In many cases a pair of arguments may fail to be distinctly comparable in strength; neither is distinctly stronger than the other, nor are they determinately equal in strength. Nevertheless, I will argue that each legitimate *comparative support relation* should be *syntactically extendable* to a *complete* relation, at least in principle. I'll provide that argument in a moment. Let's first define the relevant notion of *extendability*.

> Definition: A *proto-support relation* $\succcurlyeq_\alpha$ is *extendable* to a *proto-support relation* $\succcurlyeq_\beta$ just when the language of $\succcurlyeq_\beta$ contains the language of $\succcurlyeq_\alpha$ (i.e. contains the same syntactic expressions, and perhaps additional expressions as well) and the following two conditions hold:

[11] Notice that rule 9 does not presume that any such lotteries exist. It only supposes that we can construct arguments describing them and their implications for prospective outcomes.

(1) whenever A|B $\succ_\alpha$ C|D, then also A|B $\succ_\beta$ C|D;
(2) whenever A|B $\approx_\alpha$ C|D, then also A|B $\approx_\beta$ C|D.

When $\succcurlyeq_\alpha$ is *extendable* to $\succcurlyeq_\beta$, all argument pairs that are *distinctly comparable* according to $\succcurlyeq_\alpha$ must *compare in the same way* according to $\succcurlyeq_\beta$. Each relation $\succcurlyeq_\alpha$ counts as an extension of itself. An extension $\succcurlyeq_\beta$ of $\succcurlyeq_\alpha$ may employ precisely the same language as $\succcurlyeq_\alpha$, and may merely extend $\succcurlyeq_\alpha$ by *distinctly comparing* some arguments that were *not distinctly comparable* according to $\succcurlyeq_\alpha$. More generally, $\succcurlyeq_\beta$ may include comparisons that involve new expressions, not already part of the syntax of the language for $\succcurlyeq_\alpha$. Furthermore, the relationship between $\succcurlyeq_\alpha$ and an *extension* of it, $\succcurlyeq_\beta$, need only be syntactic. There is no presumption that an *extension* of a relation must maintain the same meanings (the same semantic content) for sentences it shares with the relation it extends (although it certainly may do so).

The *proto-support relations* commonly permit a wide range of argument pairs to remain *incomparable in strength*. But only those relations among them that can be *extended* to *complete relations* (i.e. which satisfy rule 8) will be counted among the full-fledged *comparative support relation*. To see that *extendability* to a relation that satisfies the *completeness rule* is a plausible constraint, let's consider what a *proto-support relation* must be like if it cannot possibly be extended to a *complete relation*.

Extendability is a purely syntactic requirement. That is, an *extension* $\succcurlyeq_\beta$ of a relation $\succcurlyeq_\alpha$ need not take on any of the meanings that one might have associated with the sentences of $\succcurlyeq_\alpha$. Rather, an *extension* $\succcurlyeq_\beta$ of $\succcurlyeq_\alpha$ is required only to agree with the *definite comparisons*—those of form A|B $\succ_\alpha$ C|D and E|F $\approx_\alpha$ G|H already specified by $\succcurlyeq_\alpha$ (while $\succcurlyeq_\beta$ continues to satisfy axioms 0-7). Thus, a proto-support relation $\succcurlyeq_\alpha$ may fail to be extendable to a complete relation only when no complete extension of $\succcurlyeq_\alpha$ is consistent with the (purely syntactic) restrictions on orderings embodied by axioms 0–7. That is, for $\succcurlyeq_\alpha$ to be *unextendable* to a complete relation, the definite comparisons (of form A|B $\succ_\alpha$ C|D and E|F $\approx_\alpha$ G|H) already specified by $\succcurlyeq_\alpha$ must, in conjunction with axioms 0–7, *require* that some argument pairs inevitably remain incomparable in strength *merely to avoid an explicit syntactic contradiction*. In other words, any proto-support relation for which there cannot possibly be a *complete extension* must already contain a kind of looming syntactic inconsistency, owing to the forms of its *definite* argument strength comparisons. It manages to stave off explicit formal inconsistency only by forcing at least some argument pairs to remain incomparable in strength.

It makes good sense to declare specific argument pairs *incomparable in strength* when, given their meanings (their semantic content), there seems to be no appropriate basis on which to compare them. But let's disregard any comparative relation that *requires* argument forms to remain *incomparable* in order to avoid syntactic inconsistency. Thus, we disregard any relation that cannot possibly be extended to a *complete relation*, not even by radically changing the meanings of the sentences involved.

We now define the class of *comparative support relations* as those *proto-support relations* that can be extended to relations that satisfy rules 8 and 9. Those that satisfy rules 8 and 9^+ are a special subclass of *comparative support relations*, the *arch* (for "Archimedean") *comparative support relations*.

Definition: Classes of Comparative Support Relations:

1. A ***completely-extended*** *proto-support relation* is any *proto-support relation* that satisfies rules 8 and 9.
2. A ***completely-extendable*** *proto-support relation* is any *proto-support relation* that is *extendable* to a ***completely-extended*** *proto-support relation* (i.e. any *proto-support relation* is extendable to a *proto-support relation* that satisfies rules 8 and 9).

Define the *comparative support relations* to be the ***completely-extendable*** *proto-support relations*. They include the ***completely-extended*** relations as special cases.

3. An ***completely-arch-extended*** *proto-support relation* is any *proto-support relation* that satisfies rules 8 and 9^+.
4. A ***completely-arch-extendable*** *proto-support relation* is any *proto-support relation* that is *extendable* to a ***completely-extended*** *proto-support relation* (i.e. any *proto-support relation* is extendable to a *proto-support relation* that satisfies rules 8 and 9^+).

Define the *arch comparative support relations* to be the ***completely-arch-extendable*** *proto-support relations*. They include the ***completely-arch-extended*** relations as special cases.

When a *proto-support relation* $\succcurlyeq_\alpha$ is extendable to one that satisfies rule 9 (together with 8) but *not* to one that satisfies 9^+, for some argument pairs, although $A|B \succ_\alpha C|D$, the relation *requires* that A|B be *only infinitesimally stronger* than C|D. That is, the relationships among arguments already specified by $\succcurlyeq_\alpha$ must imply that no extension of it, $\succcurlyeq_\beta$, can permit an equal-partition argument, $(S_1 \vee \ldots \vee S_m)|F$ to fit between A|B and C|D (for any n-sized equal-partition, where $F \Rightarrow_\beta \neg(S_i \cdot S_j)$, $S_i|F \approx_\beta S_j|F$, $F \Rightarrow_\beta (S_1 \vee \ldots \vee S_n)$). Such *non-Archimedean support relations* turn out to have interesting features, so we won't entirely bypass them here.

Each *completely-extendable relation* is representable by a Popper function—perhaps by more than one. A typically *completely-extendable* relation may be extended to a variety of distinct *completely-extended* relations. Each *completely-extended relation* is represented by a unique Popper function. The nature of this probabilistic representation is perfectly tight for the arch relations, and a bit looser for the other relations. Here are the precise details.

Representation Theorem for *comparative support relations*:

1. For each *completely-arch-extended comparative support relation* $\succcurlyeq$, there is a unique Popper function P such that for all sentence H_1, E_1, H_2, E_2 in the language of $\succcurlyeq$, $P[H_1|E_1] \geq P[H_2|E_2]$ if and only if $H_1|E_1 \succcurlyeq H_2|E_2$.
   Note 1: given that $\succcurlyeq$ is *completely-arch-extended*, this condition is equivalent to the conjunction of the following:
   (1) if $P[H_1|E_1] > P[H_2|E_2]$, then $H_1|E_1 \succ H_2|E_2$;
   (2) if $P[H_1|E_1] = P[H_2|E_2]$, then $H_1|E_1 \approx H_2|E_2$.
   Note 2: given that $\succcurlyeq$ is *completely-arch-extended*, this condition is also equivalent to the conjunction of the following:
   (1) if $H_1|E_1 \succ H_2|E_2$, then $P[H_1|E_1] > P[H_2|E_2]$;
   (2) if $H_1|E_1 \approx H2|E_2$, then $P[H_1|E_1] = P[H_2|E_2]$.
2. For each *arch comparative support relation* $\succcurlyeq$ (which, by definition, must be a *completely-arch-extendable proto-support relation*), there is a (not necessarily

unique) Popper function P such that for all sentence H_1, E_1, H_2, E_2 in the language of $\succcurlyeq$,
if *not* $H_1|E_1 \asymp H_2|E_2$, then $P[H_1|E_1] \geq P[H_2|E_2]$ if and only if $H_1|E_1 \succcurlyeq H_2|E_2$.
Note 1: this condition is equivalent to the conjunction of the following:
(1) if $P[H_1|E_1] > P[H_2|E_2]$, then $H_1|E_1 \succ H_2|E_2$ or $H_1|E_1 \asymp H_2|E_2$;
(2) if $P[H_1|E_1] = P[H_2|E_2]$, then $H_1|E_1 \approx H_2|E_2$ or $H_1|E_1 \asymp H_2|E_2$.
Note 2: this condition is also equivalent to the conjunction of the:
(1) if $H_1|E_1 \succ H_2|E_2$, then $P[H_1|E_1] > P[H_2|E_2]$;
(2) if $H_1|E_1 \approx H_2|E_2$, then $P[H_1\ |E_1] = P[H_2|E_2]$.

3. For each *completely-extended comparative support relation* $\succcurlyeq$, there is a unique Popper function P such that for all sentence H_1, E_1, H_2, E_2 in the language of $\succcurlyeq$,
if $P[H_1|E_1] > P[H_2|E_2]$, then $H_1|E_1 \succ H_2|E_2$.
Note: given the *completeness* of , this condition is equivalent to the conjunction of the following conditions:
(1) if $H_1|E_1 \succ H_2|E_2$, then $P[H_1|E_1] \geq P[H_2|E_2]$;
(2) if $H_1|E_1 \approx H_2|E_2$, then $P[H_1\ |E_1] = P[H_2|E_2]$.
4. For each *comparative support relation* $\succcurlyeq$ (which, by definition, must be a *completely-extendable proto-support relation*), there is a (not necessarily unique) Popper function P such that for all sentence H_1, E_1, H_2, E_2 in the language of $\succcurlyeq$,
if $P[H_1|E_1] > P[H_2|E_2]$, then $H_1|E_1 \succ H_2|E_2$ or $H_1|E_1 \asymp H_2|E_2$.
Note: this condition is equivalent to the following pair of conditions:
(1) if $H_1|E_1 \succ H_2|E_2$, then $P[H_1|E_1] \geq P[H_2|E_2]$;
(2) if $H_1|E_1$ " $H_2|E_2$, then $P[H_1\ |E_1] = P[H_2|E_2]$.

The representation theorem shows that each *completely-arch-extended* relation is virtually identical to its uniquely representing Popper function. Whenever the representing *Popper function* P for ***complete-arch*** *relation* $\succcurlyeq$ assigns $P[A\ |B] = r$ for a *rational* number $r = m/n$, the relation $\succcurlyeq$ acts like the representing probability function via a rule-9^+-style-satisfying partition for which $A|B \approx (S_1 \vee ... \vee S_m)|(G \cdot (S_1 \vee ... \vee S_n))$.[12] When the representing function for $\succcurlyeq$ assigns $P[A\ |B] = r$ for an *irrational* number r, the relation $\succcurlyeq$ supplies a sequence of increasingly large rule-9^+-style partitions (for ever larger n), which supply a sequence of relationships $(S_1 \vee ... \vee S_{m+1})|(G \cdot (S_1 \vee ... \vee S_n)) \succcurlyeq A|B \succcurlyeq (S_1 \vee ... \vee S_m)|(G \cdot (S_1 \vee ... \vee S_n))$, where the values of m/n and (m+1)/n converge to r as n increases. Associated with each *completely-arch-extendable relation* is the set of Popper functions that represent its various *complete-arch extensions*.[13]

More generally, each *completely-extended* relation (*arch* or not) is at least nearly identical to its uniquely representing Popper function. A completely-extended relation $\succcurlyeq$ that fails to be *completely-arch-extended* may exhibit a slight "looseness in fit" of the following kind: its representing Popper function P may assign $P[A\ |B] = P[C|D]$ in cases where

[12] I.e. for which $(S_1 \vee ... \vee S_m)|(G \cdot (S_1 \vee ... \vee S_n)) \succcurlyeq A|B \succcurlyeq (S_1 \vee ... \vee S_m)|(G \cdot (S_1 \vee ... \vee S_n))$.

[13] This is the essence of how the representation theorem is proved. The version of the proof provided by Koopman (1940) is easily adapted to the representation of *completely-extended* comparative support relations by Popper functions—clause 3 of the above theorem. The remaining clauses of this theorem follow easily from clause 3.

although A|B $\succ$ C|D, no equal-partition can be fitted between A|B and C|D (i.e. they are infinitesimally close together).

Associated with each *completely-extendable relation* is the collection of Popper functions that represent its various *complete extensions*. Each Popper function in the representing collection for $\succcurlyeq$ accurately preserves the precise orderings ($\succ$ and $\approx$) among the argument pairs it compares—except for infinitesimally close argument pairs, which are always represented as equal.[14]

13.5 Conclusion

Bayesian approaches to confirmation theory represent evidential support for hypotheses in terms of conditional probability functions, which assign precise numerical values to each argument A|B. In most cases these probability assignments are overly precise. For, in most real scientific contexts the strengths of plausibility arguments for various alternative hypotheses, as represented by prior probability assignments, are fairly indefinite, and so not properly rendered by the kinds of precise numbers that conditional probability functions assign. Indeed, most plausibility arguments for hypotheses are better rendered in terms of their strength compared with arguments for alternative hypotheses, rather than in terms of precise probability values. Thus, Bayesian prior probabilities are best represented by ranges of numerical values that capture the imprecision in comparative assessments of the extent to which plausibility considerations support one hypothesis over another.

Furthermore, in many cases this problem of over-precision also plagues the assignment of Bayesian *likelihoods* for evidence claims on various hypotheses (which represent what hypotheses *say* about the evidence). In realistic cases these will often have rather vague values. Indeed, on Bayesian accounts, the import of evidence via likelihoods is completely captured by *ratios of likelihoods*, which represent how much more likely the evidence is according to one hypothesis than according to an alternative. In many realistic cases these likelihood comparisons will also be somewhat vague or imprecise—best represented by ranges of values.[15]

[14] Classical probability functions are, in effect, just the one-level Popper functions, those where P[A|B] = 1 for all A whenever P[B|C$\vee\neg$C] = 0. The *comparative support relations* that are represented by strictly classical probability functions are just those that satisfy the following additional rule (which provides an additional restriction on the *proto-support relations*):

If $\neg$B|(C$\vee\neg$C) $\succcurlyeq$ D|D, then A|B $\succcurlyeq$ D | D for all A (i.e. if (C$\vee\neg$C) $\Rightarrow \neg$B, then B $\Rightarrow$ A).

[15] Examples: (1) Precisely how likely is the observed fit between Africa and South America if the Continental Drift hypothesis is true? The value of this likelihood is presumably rather vague. (2) To assess how the evidence supports the Drift Hypothesis, H_D, over the alternative Contractionist Hypothesis (that the continents have remained in place since the early molten Earth cooled and contracted), H_C, we need only assess *how much more likely* the evidence is according to Drift as compared to Contraction, P[E|H_D·B] / P[E | H_C·B] (where B consists of relevant background and auxiliaries). Even this likelihood comparison will be somewhat imprecise, best represented by some range of numbers that captures the vagueness of the comparison. For confirmational purposes the size of P[E|H_D·B] / P[E |H_C·B] (very large, or extremely small) is all that matters. The resulting Bayesian assessment compares posterior

This situation places the Bayesian approach to confirmation theory in an embarrassing predicament. The Bayesian approach first proposes a probabilistic logic that assigns overly precise numerical values to all arguments A|B, and then backs off from this over-precision by acknowledging that in many realistic applications the proper representation of evidential support should employ whole collections of precise probability functions that cover ranges of reasonable values for the prior probabilities of hypotheses, and also ranges for likelihoods in many cases.

The qualitative logic of comparative support described here offers a rationale for this kind of Bayesian approach to confirmation, including the introduction of classes of confirmation functions to represent ranges of values. The overly precise *probabilistic confirmation functions* are mere representational stand-ins for a deeper qualitative logic of comparative argument strength, captured by the *comparative support relations*. Each representing probability function reiterates the qualitative comparisons of argument strength endorsed by the *comparative support relation* it represents. Furthermore, the underlying qualitative logic can directly express the *incomparability in strength* among argument pairs that is common among real arguments. Although each representing probability function for a given relation $\succcurlyeq$ fills in with precise comparisons among all argument pairs, the whole collection of representing probability functions for $\succcurlyeq$ captures the incomparability in terms of the available range of ways in which comparisons could, in principle, be filled in, given the definite comparisons provided by $\succcurlyeq$.[16]

Thus, a probabilistic Bayesian confirmation theory employs overly precise *probabilistic representations* because they are computationally easier to work with than the *comparative support relations* they represent. Nevertheless, the features of evidential support that we really care about are captured by the comparative relationships among argument strengths, realized by the *comparative support relations* and their logic. The probabilistic representation of this logic merely provides a felicitous way to represent the deeper *qualitative logic of comparative support*.

Acknowledgements

I presented an early version of this paper at the *Workshop on Conditionals, Counterfactuals, and Causes in Uncertain Environments* at the University of Dusseldorf (May 20–22, 2011). Thanks to Gerhard Schurz and Matthias Unterhuber for making this workshop work so well. Thanks to all of the workshop participants for their extremely helpful questions and

probabilities on the basis of comparisons of prior plausibility, $P[H_D|B]$ / $P[H_C|B]$ (where B contains plausibility considerations) together with evidential likelihood ratios:

$P[H_D|E \cdot B] / P[H_C|E \cdot B] = (P[E|H_D \cdot B] / P[E|H_C \cdot B]) \times (P[H_D|B] / P[H_C|B])$.

Each of these ratios may best be represented by a range of values, characterized by the collection of probability functions that provide values within the appropriate ranges.

[16] Transitivity (i.e. axiom o) yields the following constraint on incomparable argument pairs:

Given any transitive relation $\succcurlyeq$, whenever $H_2|E_2 \succcurlyeq H_1|E_1$ (i.e. whenever *not* $H_2|E_2 \succcurlyeq H_1|E_1$ and *not* $H_1|E_1 \succcurlyeq H_2|E_2$), every argument H|E that is distinctly comparable to both $H_1 \mid E_1$ and $H_2|E_2$ (i.e. $H|E \succ H_1 \mid E_1$, $H|E \succ H_2|E_2$) is ether distinctly stronger than both (i.e. $H|E \succ H_1|E_1$ and $H|E \succ H_2|E_2$) or is distinctly weaker than both (i.e. $H_1|E_1 \succ H|E$ and $H_2|E_2 \succ H|E$).

feedback. I owe a special debt to John Cusvert, David Makinson, Alan Hájek, and Chris Hitchcock. Each provided comments on evolving drafts, resulting in major revisions and improvements.

References

Cozman, F. (2016) Imprecise Probabilities. (This volume, Chapter 14).

de Finetti, B. (1937/1964) La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré*. 17. pp. 1–68. [English translation in Kyburg, H. E. and Smokler, H. E. (eds.) *Studies in Subjective Probability*. New York, NY: Wiley.]

de Finetti, B. (1974) *Theory of Probability*. Vols. 1 and 2. English translation by A. Machí and A. Smith. New York, NY: Wiley.

Easwaran, K. (2016) Conditional Probability. (This volume, Chapter 9).

Field, H. (1977) Logic, Meaning, and Conceptual Role. *Journal of Philosophy*. 74. pp. 379–409.

Hawthorne, James (2014) A Primer on *Rational Consequence Relations*, *Popper Functions*, and their Ranked Structures. *Studia Logica*. 102. 4. pp. 731–49.

Keynes, J. M. (1921) *Treatise on Probability*. London: Macmillan.

Koopman, B. O. (1940) The Axioms and Algebra of Intuitive Probability. *Annals of Mathematics*. 41. pp. 269–92.

Popper, K. R. (1959) *The Logic of Scientific Discovery*. London: Hutchinson.

Savage, L. J. (1954) *The Foundations of Statistics*. New York, NY: Wiley. [2nd edition, 1972. New York, NY: Dover.]

APPENDIX

From the Sparse Axioms 1-6 for Popper functions, we derive the Robust Axioms (1)–(6), as follows.

Notice that (6) is identical to 6, and (4) follows easily from 4; so we only need derive (1)–(3) and (5). We derive useful intermediate results along the way.

(3*) if B|= A and A|= B, then P[A|C] = P[B|C]: directly from 3.

(3') if B|= A, then P[A|B] = P[C|C]: suppose B|= A; in 3 replace 'C' with 'B' to get P[A|B] ≥ P[B|B]; use 2 twice to get P[C|C] ≥ P[A|B] ≥ P[B|B] ≥ P[C|C].

(2*) P[C|C] = 1: from ax. 1, for some A and B, P[A | B] ≠ 0; from (3*) and 6, P[A|B] = P[A · A | B] = P[A|A · B] ×P[A|B], so P[A | A · B] = 1; since A · B|= A, from (3') we have 1 = P[A|A · B] = P[C|C].

(3) if B|= A, then P[A|B] = 1: by (3') and (2*).

(2') 1 ≥ P[A|B]: from (2*) and 2.

(2#) P[A|B] ≥ 0: if P[A|C] < 0, then P[A|C] ≠ 1 = P[C|C], so 5 yields P[¬A |C] = P[C|C] − P[A|C] = 1 − P[A | C] > 1 contradicting (2').

(2) 1 ≥ P[A|B] ≥ 0: by (2') and (2#).

(5*) P[A|C] + P[¬A|C] = 1 or else P[D |C] = 1 for all D: by 5 and (2*).

(1) if|= ¬A and|= B, then P[A|B] = 0: suppose|= ¬A and|= B; from (5*) and (3), P[A|B] = 1 − P[¬A|B] = 1 − 1 = 0 *(done) unless* P[D|B] = 1 for all D; but "P[D |B] = 1 for all D" when|= B contradicts axiom 1, since it would have P[E|F] = 1 for all E and F, as follows: 1

$= P[E \cdot F|B] = P[E|F \cdot B] \times P[F|B] = P[E|F \cdot B] \times 1 = P[E|F]$ (by 6 and (4), since $B \cdot F$ is logically equivalent to F when $| = B$).

(5) if $C| = \neg(A \cdot B)$, then either $P[(A \vee B)|C] = P[A|C] + P[B|C]$ or $P[D|C] = 1$ for all D:

Suppose $C| = \neg(A \cdot B)$ but *not* "$P[D|C] = 1$ for all D".

First we derive the following useful intermediate results:

$P[A \cdot B|C] = 0$, $P[B \cdot \neg B|C] = 0$, $P[A \cdot \neg A|C] = 0$, $P[\neg B|A \cdot C] = 1$, and $P[\neg A|B \cdot C] = 1$:

Since $C| = \neg(A \cdot B)$, by (3) and (5*), $P[\neg(A \cdot B)|C] = 1 = P[(A \cdot B)|C] + P[\neg(A \cdot B)|C]$, so $P[A \cdot B|C] = 0$. Since $C| = (B \vee \neg B)$, by (3) and (5*), $P[(B \vee \neg B)|C] = 1 = P[(B \vee \neg B)|C] + P[\neg(B \vee \neg B)|C]$, so $0 = P[\neg(B \vee \neg B)|C] = P[B \cdot \neg B|C]$, by (3*). Similarly, $P[A \cdot \neg A|C] = 0$. Also, from $C| = \neg(A \cdot B)$ we have both $(A \cdot C)| = \neg B$ and $(B \cdot C)| = \neg A$, so from (3) $P[\neg B|A \cdot C] = 1$ and $P[\neg A|B \cdot C] = 1$.

Now we consider three cases, showing in each we get $P[(A \vee B)|C] = P[A|C] + P[B|C]$.

Case 1: Suppose $P[B|\neg B \cdot C] = 1$: $0 = P[B \cdot \neg B|C] = P[B|\neg B \cdot C] \times P[\neg B|C] = P[\neg B|C]$, so $P[\neg B|C] = 0$, so $P[B|C] = 1$, by 6 and (5*). Since $\neg(A \vee B)| = \neg B$, $0 = P[\neg B|C] \geq P[\neg(A \vee B)|C] \geq 0$ by 3 and (2), so $P[\neg(A \vee B)|C] = 0$, then by (5*), $P[(A \vee B)|C] = 1 - P[\neg(A \vee B)|C] = 1 = P[B|C]$; thus, $P[A \vee B|C] = P[B|C]$. Also, $1 = P[\neg B|A \cdot C]$, so from $\neg B \cdot A| = \neg B$, 3 and 6, $0 = P[\neg B|C] \geq P[\neg B \cdot A|C] = P[\neg B|A \cdot C] \times P[A|C] = P[A|C]$, so $P[A|C] = 0$. Thus $P[(A \vee B)|C] = P[A|C] + P[B|C]$.

Case 2: Suppose $P[A|\neg A \cdot C] = 1$: Then (as in Case 1) $P[(A \vee B)|C] = P[A|C] + P[B|C]$.

Case 3: Suppose $P[B|(\neg B \cdot C)] < 1$ and $P[A|(\neg A \cdot C)] < 1$. Then by repeated instances of (5*), (3*) and 6: $1 - P[(A \vee B)|C] = P[\neg(A \vee B)|C] = P[\neg A \cdot \neg B|C] = P[\neg A|\neg B \cdot C] \times P[\neg B|C] = (1 - P[A|\neg B \cdot C]) \times P[\neg B|C] = P[\neg B|C] - P[A \cdot \neg B|C] = P[\neg B|C] - P[\neg B \cdot A|C] = P[\neg B|C] - P[\neg B|A \cdot C] \times P[A|C] = P[\neg B|C] - (1 - P[B|A \cdot C]) \times P[A|C] = P[\neg B|C] - P[A|C] + P[B \cdot A|C] = 1 - P[B|C] - P[A|C] + 0$; so $1 - P[(A \vee B)|C] = 1 - P[B|C] - P[A|C]$; thus, $P[(A \vee B)|C] = P[B|C] + P[A|C]$.

CHAPTER 14

IMPRECISE AND INDETERMINATE PROBABILITIES

FABIO G. COZMAN

Examples in probability theory often employ probability values that are assumed to be of utmost precision; for instance, the probability that a fair coin yields heads is exactly 0.5. However, it is hard to accept that by looking at the sky one can conclude that the probability of rain is *exactly* 0.7. Why not 0.689438? Indeed there are many situations where one may feel uncomfortable with a sharp probability value. For instance, a particular set-up may be incompletely or vaguely described; or the set-up may be described in abstracted terms, perhaps because insufficient resources are available for a more detailed description; or experts may disagree on probability values, and no single value can accommodate all opinions; or one may want to represent preferences that are only partially ordered though a set of probability values. It is also possible that one may be interested in the robustness of decisions against some unspecified features of the set-up that may appear as perturbations to probability values.

An example: An urn contains 30 red balls and 60 other balls that may be either yellow or black. A person takes a ball from the urn. One may argue that the probability that this ball is yellow is not precisely known, as it depends on the urn composition: the probability is zero if all balls are either red or black; the probability is 2/3 if all balls are red or yellow; values between 0 and 2/3 can be produced by different urn compositions. This example is related to Ellsberg's paradox, discussed in more detail later.

It seems reasonable to make a distinction between two situations. First: each event is associated with a sharp probability value, but this value is at least partially hidden because of imperfect elicitation. Sometimes this situation is said to be one where probability values are *imprecise*. A second scenario ensues when events are not associated with a sharp probability value to begin with, but are instead associated with *indeterminate probabilities* (intervals, bounds, sets).

Levi (1985) emphasizes the distinction between imprecision and indeterminacy in the context of subjective probabilities. Many authors simply conflate imprecision and indeterminacy. For instance, Walley (1991) uses the expression "imprecise probability" to encompass both cases, but reserves the expression "sensitivity interpretation" to indicate

imprecision. This dual use of the expression "imprecise probability" seems to be adopted in most current literature (Augustin et al. 2014).

We may also debate whether imprecise/indeterminate probabilities should be interpreted in aleatory or epistemic ways. For instance, one may focus on *aleatory imprecise* probabilities, where one obtains partial data about a chance set-up; or on *aleatory indeterminate* probabilities, where one deals with empirical processes that defy summarization by a single probability (Fierens and Fine 2003, Fine 1988). One may wish to represent an incompletely specified statistical model by using *epistemic imprecise* prior probabilities; or one may employ *epistemic indeterminate* probabilities to represent partially ordered preferences.

Apart from interpretation issues, there are several mathematical formalisms to represent imprecise/indeterminate probability (Augustin et al. 2014, Troffaes and De Cooman 2014): interval, lower and upper probabilities, Choquet capacities, random sets, lower and upper expectations (and previsions), and sets of probability distributions. Section 14.2 comprises a review of basic mathematical models. Yet another way to classify theories of imprecise and indeterminate probability focuses on decision-making: What should decision-makers do when they face a decision? Section 14.3 reviews basic facts about decision-making.

14.1 A Bit of History

We already find discussion of indeterminate probabilities in Jakob Bernoulli's and Lambert's analysis of arguments through probability bounds in the 18th century (Hailperin 1996, Shafer 1982). Later, the work by Boole (1854) offers interesting examples of bounds within which "it is certain that the probability sought must lie." For instance:

> The probability that it thunders upon a given day is p, the probability that it both thunders and hails is q, but of the connexion of the two phenomena of thunder and hail, nothing further is supposed to be known. Required the probability that it hails on the proposed day. [Boole obtains the interval $[q, 1 + q - p]$.]

Still in the 19th century, we find clear concern with indeterminacy in probability in von Kries' work (Fioretti 2001).

A digression: Note that since the 19th century thinkers who emphasize the aleatory interpretation of probabilities (and later frequentist statisticians) have argued against prior probabilities; however, the mere rejection of the idea that specific probabilities have empirical content is downplayed in this review. Note also that an extensive literature on probability bounds has been produced during centuries of careful work; however, mere interest in probability bounds does not imply focus on indeterminate probabilities, and as such it is also downplayed in this review.

Keynes' (1921) seminal work on probabilities seems to be the first influential defense of indeterminate probabilities. Even though Keynes did not propose a formal theory of indeterminate probabilities, he argued that some probabilities are imprecisely known, while others are really indeterminate (Keynes 1921, chapters III, XV, XVIII). At the same time a book by Knight (1921) emphasized the distinction between precisely specified probabilities (representing *risk*) and indeterminate probabilities (representing *uncertainty*):

> The practical difference between the two categories, risk and uncertainty, is that in the former the distribution of the outcome in a group of instances is known (either through calculation *a priori* or from statistics of past experience), while in the case of uncertainty this is not true, the reason being in general that it is impossible to form a group of instances, because the situation dealt with is in a high degree unique.

Knight's proposals concerning risk and uncertainty did not have an immediate impact, but indeterminate probabilities repeatedly surfaced in economics during the forties and fifties (Hurwicz 1951, Shackle 1955).

The next important landmark is a pair of papers published by Koopman (1940a,b) with the first explicitly formalized theory of lower and upper probabilities. He axiomatized the relation $b/k \succ a/h$, understood as "a on the presumption h is no more probable than b on the presumption k." From these qualitative comparisons, Koopman obtained numerical lower and upper probabilities, and finally produced sharp probabilities. Note that comparative probabilities had been proposed before, in seminal work by de Finetti (1931), and even before in work by Bernstein (1917). Since then the study of comparative relations such as "more probable than" has generated considerable literature (Fine 1973, Fishburn 1986, Hawthorne (13) this volume, Kaplan and Fine 1977).

Several important pieces appeared between 1959 and 1962. In 1960 Good presented a theory based on a *black box* that contains precise probabilities, but that receives inputs and produces outputs that may be lower and upper conditional probabilities (Good 1962). Good's theory was later detailed in a number of publications dealing with the limitations of Bayesian procedures and compromises around them (Good 1983). Also in 1960, Ellsberg delivered a paper on "ambiguity" that centres on indeterminacy of probabilities (Ellsberg 1961). An interesting example, usually referred to as "Ellsberg's paradox," is based on the urn described in the second paragraph of this chapter. Suppose one is given the urn, and one ball is to be drawn. The following options are available, where the numbers indicate the amount to be received in case the ball is of the indicated color:

Option	Red	Black	Yellow
I	100	0	0
II	0	100	0
III	100	0	100
IV	0	100	100

Most people strictly prefer I to II, and IV to III, but there cannot be a single probability distribution over urn compositions that yields such preferences if comparisons are based on expected gain. Clearly the crux of the matter is that probabilities are indeterminate (see Section 14.3).

In 1961 Kyburg presented a theory connecting frequencies with interval-valued probabilities, after preliminary publications (Kyburg Jr. 1959a,b, 1961). Kyburg's philosophy of science evolved in the following decades, but always with interval-valued probabilities (Kyburg Jr. 1974).

Another remarkable development was Smith's theory of "medial odds" (Smith 1961), where the focus is on subjective preferences expressed through betting of the Ramsey/de Finetti variety (Zynda (17) this volume). To understand the idea, consider a bet on event A

that pays 1 if A occurs and 0 otherwise; suppose the decision-maker buys the bet for any price strictly smaller than p, and sells the bet for any price strictly greater than p. Following de Finetti, p is understood as the probability of A. Now it is clearly unrealistic to require that buying and selling prices have a common extreme value; Smith's much more realistic idea is to suppose that one will buy for any price strictly smaller than $\underline{p}$, and sell for any price strictly greater than $\bar{p}$, where $\underline{p}$ may be strictly smaller than $\bar{p}$. Smith takes these values respectively as the lower and the upper probabilities of A, and any value between $\underline{p}$ and $\bar{p}$ is a "medial probability."

In closing the period 1959–62, we note the influential defense of robustness analysis in statistics by Tukey (1960), and a derivation of utility theory without completeness by Aumann (1962), with techniques later used to axiomatize sets of probabilities (for instance by Fishburn (1970)).

The work of Huber (1964) generated frequentist statistical procedures whose robustness was investigated using sets of probability distributions; later the theory of frequentist robustness adopted several models based on Choquet capacities (Huber 1980). Already in the fifties and sixties one can find discussion of *Bayesian robust statistics*, with a focus on imprecision/indeterminacy in prior distributions (Blum and Rosenblatt 1967, Hodges and Lehmann 1952, Menges 1966), but it does not seem that the field grew up from a single seminal proposal. During the seventies and eighties Bayesian robustness received increased attention, as indicated in several reviews with substantial historical background (Berger 1984, 1990, 1994, Insua and Ruggeri 2000). Bayesian robust statistics tends to assume that there must be a single prior distribution hidden somewhere (perhaps in Good's black box), but this distribution is only partially specified.

Returning to the sixties, we find relevant pieces on incompletely specified statistical and decision problems (Fishburn 1964) and on logical arguments (Hailperin 1965, Suppes 1966). More importantly, we find work by Dempster (1967, 1968) on upper and lower probabilities; Dempster showed how to compute lower and upper expectations and introduced a rule, usually referred to as the *Dempster rule*, to combine several "independent" upper probabilities.

The seventies saw the development of Dempster's ideas in a number of ways, most notably in Shafer's (1976) theory of belief functions. Shafer's theory became known to researchers in artificial intelligence through the work of Barnett (1981), who offered it as a reasoning tool for expert systems. The Dempster-Shafer theory received explosive attention during the eighties and early nineties as practitioners in artificial intelligence felt the need to explore alternatives to standard probability theory (Yager et al. 2008). Across the years, the controversy around interpretations of Dempster-Shafer theory has sparked interest in imprecise/indeterminate probabilities.

The seventies also saw notable work on the axiomatization of upper and lower probabilities (Suppes 1974, 1975), on decision-making (Cannon and Kmietowicz 1974), on sequential decision-making (Satia and Lave 1970), on economic analysis of ambiguity (Yates and Zukowski 1976), on argumentation (Adams and Levine 1975). The decade also produced combinations of probability and logic, propounded particularly by philosophers of science. The resulting probabilistic logics often resorted to inequalities involving probability values, and equated the consistency of a set of sentences with the existence of a nonempty set of probability distributions that satisfy all the inequalities. An early example is Keisler's (1977) $L_{\omega P}$ logic, which introduces the "probability quantifiers" such as $(P\, x \geq r)$. Additionally, two important developments must be mentioned. First, Levi (1974)

started his investigation of *convex sets* of probability distributions, a more general model than considered previously; Levi (1980) presents a detailed philosophical examination of his ideas. Second, Williams presented a theory of indeterminate probabilities by extending de Finetti's betting scheme (Williams 1976, 2007), and a theory with similar features was sketched by Buehler (1976). Note that while de Finetti himself did not embrace indeterminacy, his theory does produce inferences expressed through intervals of probabilities, and he conceded that some imprecision (similar to that of Good's black box) might be faced in practice (de Finetti and Savage 1962). Indeed, de Finetti's work has had immense influence in theories of imprecise/indeterminate probability through his emphasis on betting, on expectations, on extensions; research closely following de Finetti's philosophical position has often touched on imprecise probabilities (Coletti and Scozzafava 2002).

We now accelerate into the eighties, a decade where many scattered results were consolidated. There are too many publications to cite individually; hence only a few influential efforts of foundational nature are reviewed. Giron and Rios (1980) offered a reasoned axiomatization of convex sets of distributions, while others proposed decision-making techniques (Gardenfors and Sahlin 1982, Potter and Anderson 1980, Wolfenson and Fine 1982). Several results on subjective versions of imprecise/indeterminate probability were produced by Walley (1981, 1982), as well as a theory of aleatory indeterminate probabilities (Walley and Fine 1982). Notable also is Fine (1988)'s effort to model physical phenomena using upper and lower probabilities. Within the literature on artificial intelligence, intense debate raged during the eighties on differences between various formalisms (including comparative and order of magnitude probabilities, belief functions, fuzzy logic). Particularly notable was Nilsson's (1986) rediscovery of probabilistic logic, in work that generated an enormous literature (Hansen and Jaumard 1996).

The eighties also witnessed the gradual development of axiomatic theories of imprecise/indeterminate probability within economics and finance. As research in psychology and behavioral economics in the sixties and seventies demonstrated the frailties of the usual expected utility theory, economists were searching for better decision theories, including theories dealing with ambiguity aversion (in Ellsberg's sense). Formalized theories appeared during the eighties: Bewley (2002) axiomatized sets of probability distributions following Knight's theory, while Schmeidler (1989) axiomatized decision-making based on Choquet capacities, and Gilboa and Schmeidler (1989) axiomatized sets of probability distributions coupled with minimax behavior. The latter two theories have had significant impact in economics (Gilboa 2004). There is also a sizeable literature on robust estimation and control, including applications in econometrics, that also displays a pragmatic take on ambiguity aversion and imprecise/indeterminate probabilities (Hansen and Sargent 2007). It should be noted that the extensive literature on ambiguity aversion in economics has evolved with understanding of, but no real collaboration with, researchers in other fields mentioned in this review, for instance robust statistics.

It is fair to say that at the beginning of the nineties there were several detailed theories dealing with imprecise/indeterminate probabilities, ranging from comparative probabilities through lower/upper probabilities to sets of probability distributions. This body of knowledge was disconnected and scattered in several fields, with applications mostly in philosophy, psychology, economics, statistics, and artificial intelligence. The publication of Walley's (1991) magisterial book on lower previsions gave interested researchers a broad and organized landscape in which to navigate. Walley's theory is an extended and modified version of Williams' (1976, 2007) earlier theory. A similar attempt at a comprehensive theory

was simultaneously published by Kuznetsov (1991), but the untimely death of the author prevented this work from having a more widespread influence.

Since then, much progress has been made. New theories have been developed, such as info-gap theory (Ben-Haim 2006), game-theoretic probability (Shafer and Vovk 2001), the theory of clouds (Neumaier 2004); gradually the similarities and differences amongst the various approaches to imprecise and indeterminate probability became understood, with the derivation of many technical results and algorithms, particularly in connection with dynamic models and structural assessments such as independence and exchangeability. In the nineties robust statistics saw a surge of interest in sets of probability distributions, together with a growing attention to foundations (Berger 1994, Insua and Ruggeri 2000). Other topics in statistics were approached, such as treatment identification (Manski 2003), model construction (Goldstein and Wooff 2007), and reliability (Coolen-Schrijner et al. 2009). Research on ambiguity aversion in economics continued to flourish, and connections with risk measures in finance were produced (Vicig 2008); even the neural basis for ambiguity aversion has been studied (Hsu et al. 2005). Artificial intelligence researchers focused on algorithms and on independence relations (Cozman 2005), and in particular classification problems received practical attention (Corani and Zaffalon 2008); additionally, probabilistic logic became a central topic that often touches on indeterminacy of probability values (Lukasiewicz 2008). These and many other applications can be observed in various fields, for instance in engineering, decision analysis, and psychology; representative papers can be found in the proceedings of meetings organized by the *Society on Imprecise Probabilities: Theories and Applications* (www.sipta.org). Two books that summarize most existing theory on imprecise/indeterminate probabilities are now available (Augustin et al. 2014, Troffaes and De Cooman 2014); these books can be consulted for detailed information.

14.2 A Bit of Modeling

The simplest assessments of probabilities are verbal ones such as "an earthquake is probable" or comparative ones such as "rain is no more probable than an earthquake." One may translate then into inequalities; for instance, $P(\text{earthquake}) \geq 1/2$. One may instead contemplate more elaborate numerical assessments, such as lower/upper or interval probabilities/expectations, Choquet capacities, belief functions, possibility measures, or inequalities specifying sets of probability distributions. We briefly sketch these alternatives in this section, noting that each one of them qualifies as a complete theory of uncertainty representation. In a sense, the focus on the terms "imprecise probability" and "indeterminate probability" is unfortunate as it suggests that such imprecision/indeterminacy is undesirable, while in fact we are interested in structures that are richer, more realistic and more general than those applied in standard probability. It would make more sense to speak explicitly of the "theory of coherent lower probability" or the "theory of credal sets".

Denote the set of possibilities or states by Ω (the sample space). To avoid unnecessary complications at this point, assume Ω is finite: events are subsets of Ω and random variables are functions from Ω to the real numbers.

14.2.1 Probability Intervals, Lower and Upper Probabilities, Belief Functions, Possibility Measures

Start with probability intervals; that is, for each event A we contemplate an interval $[\underline{P}(A), \bar{P}(A)]$, where $\underline{P}(A)$ and $\bar{P}(A)$ are respectively the lower and upper probabilities of A. What conditions should we impose on these set-functions?

One possibility is to say that lower and upper probabilities should always be generated by some set of probability measures K as follows:

$$\underline{P}(A) = \inf_{P \in K} P(A), \bar{P}(A) = \sup_{P \in K} P(A). \tag{14.1}$$

Lower probabilities that can be generated by a set K are said to be *coherent* and to be *dominated* by the probability measures in K. Note that more than a single set of probability measures may generate identical coherent probability intervals; however given coherent probability intervals, there is a single largest set of probability measures that generates them.

Now the question is whether we can axiomatize lower and upper probability directly, without any mention of dominating measures. It makes sense to impose $\underline{P}(\phi) = 0$ and $\underline{P}(\Omega) = 1$, and

$$\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) \text{ whenever } A \text{ and } B \text{ are disjoint.} \tag{14.2}$$

Additionally, upper probabilities should satisfy $\bar{p}(\varphi) = 0$ and $\bar{p}(\Omega) = 1$, and

$$\bar{P}(A \cup B) \leq \bar{P}(A) + \bar{P}(B) \text{ whenever } A \text{ and } B \text{ are disjoint,}$$

and we should have the conjugacy property $\underline{P}(A) = 1 - \overline{P}(A^c)$ where A^c denotes the complement of A.

A more compact strategy is to axiomatize *only* lower probability through $\underline{P}(\phi) = 0$, $\underline{P}(\Omega) = 1$, Expression (14.2), and

$$\underline{P}(A \cap B) + 1 \geq \underline{P}(A) + \underline{P}(B) \text{ whenever } A \text{ and } B \text{ are disjoint;}$$

all properties of upper probabilities are then obtained using the conjugacy property.

However, these conditions do not guarantee that lower probabilities are always dominated by a probability measure (Fine 1988, Fine Chapter 11 in this volume, Sadrolhefazi and Fine 1994). Even though there are many conditions that guarantee dominance and coherence of lower probabilities (Huber 1980), none seems to be considered intuitive. Besides, it is difficult to handle conditioning and independence with interval probabilities. In response to these difficulties, Weichselberger and collaborators have developed a theory of interval probabilities that is explicitly based on the set of dominating probabilities measures, called the *structure* of the lower probabilities (Weichselberger et al. 2001).

An additional, often assumed, condition on lower probabilities is:

$$\underline{P}(A \cup B) \geq \underline{P}(A) + \underline{P}(B) - \underline{P}(A \cap B) \text{ for all events } A \text{ and } B.$$

A lower probability that satisfies this condition is also a *2-monotone Choquet capacity*; such a lower probability is always coherent. The literature on applications of 2-monotone

capacities, for instance in robust statistics and ambiguity aversion, is huge (Berger 1994, Huber 1980, Insua and Ruggeri 2000, Gilboa 2004). There exist definitions for n-monotone capacities for $n > 2$; when a capacity satisfies n-monotonicity for all n, we refer to it as a *belief function* (Shafer 1976).

There are many ways to describe a belief function; here is the original one due to Dempster (1967). Suppose we have two spaces Ω and Ω', and there is a probability measure over Ω'. Now suppose there is a multivalued mapping Γ taking each element of Ω' into a set of elements in Ω. An event B in Ω cannot be associated with a single probability value, for it may "receive" mass from several elements of Ω'. If $\Gamma(.)$ is always nonempty, the lower probability of B can be written as

$$\underline{P}(B) = \sum_{\omega \in \Omega' : \Gamma(\omega) \subseteq B} P(\omega).$$

A similar process can be constructed to interpret B as a *random set*, with a vast array of applications (Molchanov 2005).

An even more specialized model is used in possibility theory: here one focuses on set-functions μ such that there is a function $\pi : \Omega \to [0,1]$ where

$$\mu(A) = sup(\pi(\omega) : \omega \in A) \quad \text{for all } A \text{ in } \Omega.$$

The function π is called the *possibility distribution* of the *possibility measure* μ. Any such μ is a upper probability with generating probability measure concentrated in nested sets.

Thus we have that possibility measures, belief functions, 2-monotone capacities, and lower probabilities are increasingly more general models (each model strictly containing the previous ones).

14.2.2 Lower and Upper Expectations, Credal Sets

Now consider a slight change of focus, achieved by paying attention to expected values. Take again a set of probability measures K and note that any random variable X can be associated with *lower* and *upper expectations*:

$$\underline{E}[X] = \inf_{P \in K} E_P[X], \bar{E}[x] = \sup_{P \in K} E_P[X] \tag{14.3}$$

(where $E_P[X]$ is the expectation of X with respect to P). We can think of a set of expectation functionals, each one of them *dominating* the lower expectation functional $\underline{E}$ for all X. Whenever a lower expectation functional $\underline{E}$ is generated by a set K of standard probability measures, it is said to be *coherent*.

Now the question is whether we can axiomatize (coherent) lower and upper expectations directly, without mentioning dominating expectation functionals. The answer is positive. If we assume the following axioms for lower expectations:

$$\underline{E}[X] \geq \inf_{\omega \in \Omega} X(\omega), \underline{E}[\alpha X] = \alpha \underline{E}[X] \text{ for } \alpha > 0, \underline{E}[X + Y] \geq \underline{E}[X] + \underline{E}[Y],$$

then $\underline{E}$ is a coherent lower expectation functional induced by a convex closed set of dominating probability measures; such a set is called a *credal set* (from Levi's *credal states*).

Note that a credal set specifies a unique coherent lower expectation through Expression (3). Given coherent lower expectations, we can obtain (coherent) upper expectations through the expression $\overline{E}[X] = -\underline{E}[-X]$.

In short, it is much easier mathematically, and conceptually, to specify coherent lower expectations than to specify coherent lower probabilities. Hence from a foundational perspective it makes sense to consider expectation as the primary concept, as advocated by de Finetti (1974–5). Note that de Finetti uses the term *prevision* for (essentially) expectations; Walley refers to *coherent lower previsions* instead of coherent lower expectations, but for the purposes of this review both concepts are equivalent. The theory of lower previsions, already examined in detail by Walley (1991), has received enormous attention (Troffaes and De Cooman 2014).

One might wish to derive coherent lower expectations from even more basic axioms with supposedly intuitive content. There are two strategies for doing so: by formalizing one-sided betting, or by representing preferences.

Consider one-sided betting (Smith 1961, Walley 1991, Williams 1976). Here a random variable X is interpreted as a gamble that can be bought or sold. Now to buy X at β is to accept $X - \beta$ as profitable; and to sell X at γ is to accept $\gamma - X$ as profitable. Suppose we interpret $\underline{E}[X]$ as the supremum buying price of X, and $\overline{E}[X]$ as the infimum selling price of X. Thus one agrees to buy X at less than $\underline{E}[X]$, and to sell X at more than $\overline{E}[X]$. An intuitive condition is to avoid sure loss: for any n, we must have:[1]

$$\sup_{\omega \in \Omega} \sum_{i=1}^{n} \left(X_i(\omega) - \underline{E}[X_i] \right) \geq 0;$$

The important result is that a lower expectation functional is dominated by an expectation functional iff this condition is satisfied.

The following additional condition on one-sided betting is true iff the lower expectation is coherent: for any n, m, X_0:

$$\sup_{\omega \in \Omega} \sum_{i=1}^{n} \left(X_i(\omega) - \underline{E}[X_i] \right) - m \left(X_0(\omega) - \underline{E}[X_0] \right) \geq 0. \tag{14.4}$$

To interpret this condition, suppose it fails. Then there exists some $\delta > 0$ such that $m(X_0 - (\underline{E}[X_0] + \delta))$ is larger than a a linear combination of acceptable transactions (an acceptable joint transaction). But then $m(X_0 - (\underline{E}[X_0] + \delta))$ is acceptable and $\underline{E}[X]$ is not the supremum buying price as originally intended, and this is an inconsistency. Walley (1991) shows that given any set of lower expectation values that satisfies Expression (14.4), one can construct a minimal coherent lower expectation functional that he calls the *natural extension* of the lower expectation values. Recent developments on one-sided betting are reviewed by Miranda (2008) and by Troffaes and De Cooman (2014).

Note that if we focus on one-sided betting only with gambles that are indicator functions, we obtain an axiomatization of coherent lower probabilities (Williams 1976).

[1] There are technical subtleties concerning conditions on the space of random variables, but we avoid these here.

Axiomatization of partially ordered preferences is another way to generate a coherent lower expectation and consequently to generate a credal set (Fishburn 1970, Giron and Rios 1980, Walley 1991). For instance, postulate a preference relation amongst gambles; say that $X \succcurlyeq Y$ means that Y is not preferred to X. Suppose we assume that $\succcurlyeq$ is reflexive and transitive. Moreover, suppose $\succcurlyeq$ satisfies dominance (that is, if X is always larger than Y, then $X \succcurlyeq Y$ and not $Y \succcurlyeq X$) and some form of continuity. Finally, suppose $\succcurlyeq$ satisfies the "independence" postulate (that is, for any $\alpha \in (0, 1)$, $X \succcurlyeq Y$ iff $\alpha X + (1-\alpha)Z \succcurlyeq \alpha Y + (1-\alpha)Z$). Then $\succcurlyeq$ is represented by a credal set K:

$$X \succcurlyeq Y \text{ iff } \forall P \in K\text{: } E[X] \geq E[Y].$$

This approach can be understood as starting from an axiomatization of comparative preferences, and then adding conditions that ensure a numerical representation. Walley (2000) has advocated the study of sets of desirable gambles by themselves, without the need to explicitly consider probabilities.

A different kind of axiomatization for preferences imposes a complete ordering on $\succcurlyeq$ but relaxes the "independence" postulate (Gilboa and Schmeidler 1989, Gilboa 2004, Schmeidler 1989). In this case one obtains a representation of $\succcurlyeq$ that employs either a 2-monotone capacity or a credal set with a rule that completely orders the gambles.

There have been proposals for sets of probability distributions without closure and convexity conditions (Gardenfors and Sahlin 1982, Kyburg and Pittarelli 1996, Seidenfeld et al. 1990, 2007), and proposals that advocate sets of utility-probability pairs (Nau 2006, Seidenfeld et al. 1995, Insua 1992).

14.2.3 Conditioning, Independence, Exchangeability

Theories of imprecise/indeterminate probabilities usually offer proposals for conditioning, independence, and further structural judgments such as exchangeability. In fact, much of the debate since 1980 centers around different ways to define and analyze such concepts. The most consistent proposals deal, in one way or another, with properties of credal sets (or properties that can be reduced to properties of credal sets). For instance, the most popular way to define conditioning is to take the conditional lower expectation of X given A to be:

$$\underline{E}[x|A] = \inf_{P \in K} E_P[X|A]$$

where K is the credal set of interest and $E_P[X|A]$ is the standard conditional expectation. Now if the upper probability of A is zero (that is, if all possible probabilities for A are zero), one may leave the conditional expectation undefined, or perhaps find methods that allow it to be separately specified (Walley 1991). The case where the lower probability of A is zero, but the upper probability of A is nonzero, is more complicated: perhaps we can leave the conditional expectation undefined, or maybe we can just discard the probability measures that assign probability zero to A (Weichselberger et al. 2001). Matters become quite complex if one looks at infinite spaces, where the assumption of countable additivity is important (Walley 1991).

The definition of independence has also generated considerable debate.

The most popular definition of independence is *strong independence*. To understand it, consider a credal set K containing probability distributions for random variables X and Y. Take the set of vertices of this credal set (note: each "vertex" is a joint probability distribution). Now X and Y are strongly independent if each vertex of K satisfies standard independence. Note that this definition relies on properties of the individual sharp probabilities in the credal set.

A different concept of independence is *epistemic irrelevance*: Y is epistemically irrelevant to X if

$$\underline{E}\left[\mathrm{f(X)}\,|\,\mathrm{A(Y)}\right] = \underline{E}\left[\mathrm{f(X)}\right]$$

for every function f of X and any event A that states that Y lands in some specified set (we might say that A belongs to the algebra generated by Y; again, we avoid discussions of measurability and infinite spaces). The intuition is simple: no matter what we observe about Y, opinions about X are not changed, where opinions are represented by lower expectations.

One notable property of epistemic irrelevance is that it is asymmetric: we may have that X is epistemically irrelevant to Y and yet Y is not epistemically irrelevant to X. Hence Walley (1991, chapter 9) symmetrizes the concept: *epistemic independence* of X and Y means that X is epistemically irrelevant to Y and vice versa. There are several other possible concepts of independence; the topic has been reviewed by Cozman (2012). One important development has been the derivation of limiting laws for concepts of independence and exchangeability (De Cooman and Miranda 2008, De Cooman et al. 2009).

14.3 A Bit of Decision-making

Suppose we have a coherent lower expectation, or equivalently, a credal set K, and a set of options (each option is a random variable yielding a utility at each state of nature). How do we make decisions? There are many possibilities; the matter has been subject to intense philosophical and practical debate (Troffaes 2007, Troffaes and De Cooman 2014).

We must differentiate criteria that generate a complete order over options, and criteria that generate a partial order.

Consider first criteria that generate a complete ordering. For instance, the minimax approach used in most robust statistics (Blum and Rosenblatt 1967) and economics (Hurwicz 1951), and axiomatized by Gilboa and Schmeidler (1989) involves the choice of an option with maximum lower expectation. Take for instance Ellsberg's paradox: Options I and IV are not associated with Knightian uncertainty and have expected values of 100/3 and 200/3 respectively; Option II has a worst-case expected value of 0; Option III has a worst-case expected value of 100/3. Hence a minimaxer takes I over II and IV over III.

Other complete-ordering proposals have advocated selection of maximum upper expectation, and maximization of linear combination of lower and upper expectation.

On the other hand, axiomatizations that allow partially ordered preferences (Fishburn 1970, Giron and Rios 1980, Walley 1991) may leave the decision-maker with several incomparable non-dominated options. For instance, given a credal set, there may be a set of *maximal* options, where an option is maximal if there is no other option that dominates

it (has larger expectation value) with respect to all probability measures in the credal set (Walley 1991 chapter 3). There may also be a set of *E-admissible* options, where an option is E-admissible if it is optimal (as measured by expectation) with respect to at least one probability measure in the credal set (Levi 1980). Maximality and E-admissibility coincide when the space of options is convex, but not in general (Walley 1991 Thm. 3.9.5). Moreover, there are additional proposals that generate partial orderings; for instance, one may discard an option only if its upper expectation is smaller than the lower expectation of some other option. Selecting a single option may not be easy, and there is considerable debate on how to do it. Some critics of imprecise/indeterminate probability have argued that it is undesirable not to have a complete order over options. Whether or not this is undesirable is a difficult question; there are many practical situations where judgment must be suspended so more reasoning can be exercised. It is also notable that trading markets under ambiguity do suffer from suspension of trading (Gilboa 2004).

Sequential decision problems with indeterminate probabilities invite even more philosophical inquiry; for instance, a criterion that accommodates indeterminate probabilities but demands complete orderings over options may lead to inconsistency in sequential decision making, where a decision that was planned to be taken at some future time is actually not preferred when that time arises (Seidenfeld et al. 1990, Seidenfeld 1994). What may be the best general way to plan ahead under probabilistic indeterminacy remains an open question.

References

Adams, E. W. and Levine, H. P. (1975) On the uncertainties transmitted from premises to conclusions in deductive reasoning. *Synthese*. pp. 429–60.

Augustin, T., Coolen, F. P. A., De Cooman, G., and Troffaes, M. C. M. (eds.) (2014) *Introduction to Imprecise Probabilities*. Chichester: Wiley.

Aumann, R. J. (1962) Utility theory without the completeness axiom. *Econometrica*. 30. 3. pp. 445–62.

Barnett, J. A. (1981) Computational methods for a mathematical theory of evidence. *Conference on Artificial Intelligence*. pp. 868–75.

Ben-Haim, Y. (2006) *Info-Gap Theory: Decisions Under Severe Uncertainty.* 2nd edition. Oxford: Academic Press.

Berger, J. O. (1984) Robustness of Bayesian analysis. In Kadane, J. (ed.) *Robustness in Bayesian Statistics*. New York, NY: North-Holland.

Berger, J. O. (1990) Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*. 25. pp. 303–28.

Berger, J. O. (1994) An overview of robust Bayesian analysis (with discussion). *Test*. 3. p. 5124.

Bernstein, S. N. (1917) On the axiomatic foundation of probability theory. *Soob. Khar'k. ob-va*. 15. pp. 209–74.

Bewley, T. F. (2002) Knightian decision theory. Part I. (published 1986 as technical report 807, Yale University). *Decisions in Economics and Finance*. 25. pp. 79–110.

Blum, J. R. and Rosenblatt, J. (1967) On partial a priori information in statistical inference. *Annals of Mathematical Statistics*. 38. 6. pp. 1671–8.

Boole, G. (1854) *The Laws of Thought*. London: Walton and Maberly.

Buehler, R. J. (1976) Coherent preferences. *The Annals of Statistics*. 4. 6. pp. 1051–64.

Cannon, C. M. and Kmietowicz, Z. W. (1974) Decision theory and incomplete knowledge. *Journal of Management Studies*. 11. 3. pp. 224–32.

Coletti, G. and Scozzafava, R. (2002) *Probabilistic Logic in a Coherent Setting*. Dordrecht: Kluwer.

Coolen-Schrijner, P., Coolen, F., Troffaes, M., and Augustin, T. (2009) Special issue on statistical theory and practice with imprecision. *Journal of Statistical Theory and Practice*. 3.

Corani, G. and Zaffalon, M. (2008) Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. *Journal of Machine Learning Research*. 9. pp. 581–621.

Cozman, F. G. (2005) Graphical models for imprecise probabilities. *International Journal of Approximate Reasoning*. 39. 2–3. pp. 167–84.

Cozman, F. G. (2012) Sets of probability distributions, independence, and convexity. *Synthese*. 186. 2. pp. 577–600.

De Cooman, G. and Miranda, E. (2008) Weak and strong laws of large numbers for coherent lower previsions. *Journal of Statistical Planning and Inference*. 138. 8. pp. 2409–32.

De Cooman, G. Quaeghebeur, E., and Miranda, E. (2009) Exchangeable lower previsions. *Bernoulli*. 15. 3. pp. 721–35.

de Finetti, B. (1931) Sul significato soggettivo della probabilità. *Fundamenta Mathematicae*. 17. pp. 298–329.

de Finetti, B. (1974–5) *Theory of Probability: A Critical Introductory Treatment*. Vols. 1–2. New York, NY: Wiley.

de Finetti, B. and Savage, L. J. (1962) Sul modo di scegliere le probabilità iniziali. *Biblioteca del Metron. Ser. C*. I. 81. p. 154.

Dempster, A. P. (1967) Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*. 38. pp. 325–39.

Dempster, A. P. (1968) A generalization of Bayesian inference. *Journal of the Royal Statistical Society. B*. 30 pp. 205–47.

Ellsberg, D. (1961) Risk, ambiguity, and the Savage axioms. *The Quarterly Journal of Economics*. 75. 4. pp. 643–69.

Fierens, P. I. and Fine, T. L. (2003) Towards a chaotic probability model for frequentist probability: The univariate case. *International Symposium on Imprecise Probability and Applications*. pp. 245–59.

Fine, T. L. (1973) *Theories of Probability: an Examination of Foundations*. Oxford: Academic Press.

Fine, T. L. (1988) Lower probability models for uncertainty and nondeterministic processes. *Journal of Statistical Planning and Inference*. 20. pp. 389–411.

Fine, T. L (2016) Mathematical Alternatives to Standard Probability that Provide Selectable Degrees of Precision. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Chapter 11. Oxford: Oxford University Press.

Fioretti, G. (2001) Von Kries and the other "German Logicians": Non-numerical probabilities before Keynes. *Economics and Philosophy*. 17. pp. 245–73.

Fishburn, P. C. (1964) *Decision and Value Theory*. New York, NY: Wiley.

Fishburn, P. C. (1970) *Utility Theory for Decision Making*. New York, NY: Wiley.

Fishburn, P. C. (1986) The axioms of subjective probability. *Statistical Science*. 1. 3. pp. 335–58.

Gardenfors, P. and Sahlin, N. E. (1982) Unreliable probabilities, risk taking and decision making. *Synthese*. 53. pp. 361–86.

Gilboa, I. (2004) *Uncertainty in Economic Theory: Essays in Honor of David Schmeidler's 65th Birthday*. London: Routledge.

Gilboa, I. and Schmeidler, D. (1989) Maxmin expected utility with non-unique prior. *Journal of Mathematical Economics*. 18 pp. 141–53.

Giron, F. J. and Rios, S. (1980) Quasi-Bayesian behaviour: A more realistic approach to decision making? *Bayesian Statistics*. pp. 17–38.

Goldstein, M. and Wooff, D. A. (2007) *Bayes Linear Statistics: Theory and Methods*. Chichester: Wiley.

Good, I. J. (1962) Subjective probability as a measure of a non-measurable set. *Logic, Methodology and Philosophy of Science (1960 International Congress)*. pp. 319–29.

Good, I. J. (1983) *Good Thinking: The Foundations of Probability and its Applications*. Minneapolis, MN: University of Minnesota Press.

Hailperin, T. (1965) Best possible inequalities for the probability of a logical function of events. *American Mathematical Monthly*. 72. pp. 343–59.

Hailperin, T. (1996) *Sentential Probability Logic*. Cranbury, NJ: Lehigh University Press.

Hansen, P. and Jaumard, B. (1996) *Probabilistic Satisfiability.* Technical Report G-96-31. Les Cahiers du GERAD. École Polytechnique de Montréal.

Hansen, L. P. and Sargent, T. J. (2007) *Robustness*. Princeton, NJ: Princeton University Press.

Hawthorne, J. (2016). Probability and nonclassical logic. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Chapter 13. Oxford: Oxford University Press.

Hodges, J. L., Jr. and Lehmann, E. L. (1952) The use of previous experience in reaching statistical decisions. *Annals of Mathematical Statistics.* 23. pp. 396–407.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C. F. (2005) Neural systems responding to degrees of uncertainty in human decision-making. *Science*. 310. pp. 1680–3.

Huber, P. J. (1964) Robust estimation of a location parameter. *Annals of Mathematical Statistics.* 35. pp. 73–101.

Huber, P. J. (1980) *Robust Statistics*. New York, NY: Wiley.

Hurwicz, L. (1951) *Generalized Bayes-Minimax principle: A criterion for decision-making under uncertainty*. Technical Report Paper 355, Cowles Commission.

Insua, D. R. (1992) On the foundations of decision making under partial information. *Theory and Decision*. 33. 1. pp. 83–100.

Insua, D. R. and Ruggeri, F. (2000) *Robust Bayesian Analysis*. New York, NY: Springer Verlag).

Kaplan, M. and Fine, T. L. (1977) Joint orders in comparative probability. *The Annals of Probability.* 4. 2. pp. 161–79.

Keisler, H. J. (1977) Hyperfinite model theory. *Logic Colloquium*. 76. pp. 5–110.

Keynes, J. M. (1921) *A Treatise on Probability*. London: Macmillan and Co.

Knight, F. H. (1921) *Risk, Uncertainty and Profit*. Boston, MA: Houghton Mifflin.

Koopman, B. O. (1940a) The bases of probability. *Bulletin of the American Mathematical Society*. 46. pp. 763–74.

Koopman, B. O. (1940b) The axioms and algebra of intuitive probability. *Annals of Mathematics.* 41. 2. pp. 269–92.

Kuznetsov, V. P. (1991) *Interval Statistical Methods*. Moscow: Radio i Svyaz Publ.

Kyburg, H. E., Jr. (1959a) Probability and randomness I. *Journal of Symbolic Logic*. 24. pp. 316–17.

Kyburg, H. E., Jr. (1959b) Probability and randomness II. *Journal of Symbolic Logic*. 24. pp. 317–18.

Kyburg, H. E., Jr. (1961) *Probability and the Logic of Rational Belief*. Middletown, CT: Wesleyan University Press.

Kyburg, H. E., Jr. (1974) *The Logical Foundations of Statistical Inference*. Dordrecht: Reidel.

Kyburg, H. E., Jr. and Pittarelli, M. (1996) Set-based Bayesianism. *IEEE Trans. on Systems, Man, and Cybernetics*. A. 26. 3. pp. 324–39.

Kyburg, H. E., Jr. and Teng, C. M. (2001) *Uncertain Inference*. Cambridge: Cambridge University Press.

Levi, I. (1974) On indeterminate probabilities. *Journal of Philosophy*. 71. pp. 391–418.

Levi, I. (1980) *The Enterprise of Knowledge*. Boston, MA: MIT Press.

Levi, I. (1985) Imprecision and indeterminacy in probability judgment. *Philosophy of Science*. 52. pp. 390–409.

Lukasiewicz, T. (2008) Expressive probabilistic description logics. *Artificial Intelligence*. 172. 6–7. pp. 852–83.

Manski, C. F. (2003) *Partial Identification of Probability Distributions*. New York, NY: Springer.

Menges, G. (1966) On the Bayesification of the minimax principle. *Unternehmensforschung*. 10. pp. 81–91.

Miranda, E. (2008) A survey of the theory of coherent lower previsions. *International Journal of Approximate Reasoning*. 48. pp. 628–58.

Molchanov, I. (2005) *Theory of Random Sets*. London: Springer.

Nau, R. (2006) The shape of incomplete preferences. *The Annals of Statistics*. 34. 5. pp. 2430–48.

Neumaier, A. (2004) Clouds, fuzzy sets and probability intervals. *Reliable Computing*. 10. 4. pp. 249–72.

Nilsson, N. J. (1986) Probabilistic logic. *Artificial Intelligence*. 28. pp. 71–87.

Potter, J. M. and Anderson, B. D. O. (1980) Partial prior information and decisionmaking. *IEEE Transactions on Systems, Man and Cybernetics*. 10. 3. pp. 125–33.

Sadrolhefazi, A. and Fine, T. L. (1994) Finite-dimensional distributions and tail behavior in stationary interval-valued probability models. *The Annals of Statistics*. 22. 4. pp. 1840–70.

Satia, J. K. and Lave, R. E., Jr. (1970) Markovian decision processes with uncertain transition probabilities. *Operations Research*. 21. pp. 728–40.

Schmeidler, D. (1989) Subjective probability and expected utility without additivity. *Econometrica*. 57. pp. 571–87.

Seidenfeld, T. (1994) When normal and extensive form decisions differ. *Logic, Methodology and Philosophy of Science IX*. Amsterdam: Elsevier.

Seidenfeld, T., Schervish, M. J., and Kadane, J. B. (1990) Decisions without ordering. In *Acting and Reflecting: The Interdisciplinary Turn in Philosophy*. pp. 143–70. Dordrecht: Kluwer.

Seidenfeld, T., Schervish, M. J., and Kadane, J. B. (1995) A representation of partially ordered preferences. *Annals of Statistics*. 23. pp. 2168–217.

Seidenfeld, T., Schervish, M. J., and Kadane, J. B. (2007) Coherent choice functions under uncertainty. *International Symposium on Imprecise Probability: Theories and Applications*. pp. 385–94.

Shackle, G. L. S. (1955) *Uncertainty in Economics*. Cambridge: Cambridge University Press.

Shafer, G. (1976) *A Mathematical Theory of Evidence*. New York, NY: Princeton University Press.

Shafer, G. (1982) Non-additive probabilities in the work of Bernoulli and Lambert. *Archive for History of Exact Sciences*. 19. 4. pp. 309–70.

Shafer, G. and Vovk, V. (2001) *Probability and Finance: It's Only a Game!* New York, NY: Wiley.

Smith, C. A. B. (1961) Consistency in statistical inference and decision. *Journal of the Royal Statistical Society*. B. 23. pp. 1–25.

Suppes, P. (1966) Probabilistic inference and the concept of total evidence. In Hintikka, J. and Suppes, P. (eds.) *Aspects of Inductive Logic*. Amsterdam: North-Holland.

Suppes, P. (1974) The measurement of belief. *Journal of the Royal Statistical Society*. B. 2. pp. 160–91.

Suppes, P. (1975) Approximate probability and expectation of gambles. *Erkenntnis*. 9. pp. 153–61.

Troffaes, M. C. M. (2007) Decision making under uncertainty using imprecise probabilities. *International Journal of Approximate Reasoning*. 45. 1. pp.17–29.

Troffaes, M. C. M. and De Cooman, G. (2014) *Lower Previsions*. Chichester: Wiley.

Tukey, J. W. (1960) A survey of sampling from contaminated distributions. In Olkin, I., Ghurye, S. G., Hoeffding, W., Madow, W. G., and Mann, H. B. (eds.) *Contributions to Probability and Statistics; Essays in Honor of Harold Hotelling*. pp. 448–85. Stanford, CA: Stanford University Press.

Vicig, P. (2008) Imprecise probabilities in finance and economics. *International Journal of Approximate Reasoning*. 49. 1. pp. 99–100.

Walley, P. (1981) *Coherent Lower (and Upper) Probabilities*. Technical Report 22. Department of Statistics, University of Warwick.

Walley, P. (1982) *The Elicitation and Aggregation of Belief*. Technical Report 23. Department of Statistics, University of Warwick.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Walley, P. (2000) Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*. 24. pp. 125–48.

Walley, P. and Fine, T. L. (1982) Towards a frequentist theory of upper and lower probability. *Annals of Statistics*. 10. 3. pp. 741–61.

Weichselberger, K., Augustin, T. (assistant), and Wallner, A. (assistant) (2001) *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept*. Heidelberg: Physica-Verlag.

Williams, P. M. (1976) Indeterminate probabilities. In Przełęcki, M., Szaniawski, K., Wójcicki, R., and Malinowski, G. (eds.) *Formal Methods in the Methodology of Empirical Sciences*. pp. 229–46. Dordrecht: Reidel.

Williams, P. M. (2007) Notes on conditional previsions. (published 1975 as a technical report, School of Math. and Phys. Sci., University of Sussex). *International Journal of Approximate Reasoning*. 44. pp. 366–83.

Wolfenson, M. and Fine, T. L. (1982) Bayes-like decision making with upper and lower probabilities. *Journal of the American Statistical Association*. 77. 377. pp. 80–88.

Yager, R., Liu, L., Dempster, A. P. (advisory), and Shafer, G. (advisory) (eds.) (2008) *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Berlin, Heidelberg: Springer.

Yates, J. F. and Zukowski, L. G. (1976) Characterization of ambiguity in decision making. *Behavioral Science*. 21. pp. 19–25.

Zynda, L. (2016) Subjectivism. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Chapter 17. Oxford: Oxford University Press.

PART IV

INTERPRETATIONS AND INTERPRETIVE ISSUES

CHAPTER 15

SYMMETRY ARGUMENTS IN PROBABILITY

SANDY ZABELL

Tiger, Tiger, burning bright
In the forests of the night,
What immortal hand or eye
Could frame thy fearful symmetry?
William Blake, 'The Tiger'

Mathematical probabilities are quantitative, but the intuitions on which they are based are often qualitative. Symmetry arguments form the bridge between the two: taking us, for example, from qualitative judgments of equiprobability to quantitative statements such as "the chance of two sixes in two throws of a die is 1 in 36". Symmetry arguments, particularly in the guise of equipossible cases, dominated the early history of probability, and are still important today. This chapter both reviews this early history, and considers how such arguments have evolved in modern times.

From the birth of the probability calculus, symmetry arguments have played a central part in the subject. The use and role of those arguments, however, necessarily depends on the particular type of probability in question; at the most basic level, whether physical (an aspect of the world) versus psychological or epistemic (an aspect of knowledge); and each of these can be further subdivided into a variety of subtypes (see, e.g., Good, 1950, pp. 6–12, 1959, and 1971). In the early days of mathematical probability, however, such distinctions were neither clearly stated nor made; this necessarily complicates any historical discussion of the role of symmetry arguments. Clear statement of the difference between the two meanings first appears in the books of Poisson (1837) and Cournot (1843); and this eventually came to mark an important turning point as well in the analysis of the use of symmetry arguments in probability. In this chapter we will focus primarily on the use of symmetry in the epistemic setting rather than in the physical, for physical symmetry is largely a matter of fact rather than one of logic.

The chapter has the following structure. We begin by considering some of the presuppositions underlying modern symmetry arguments, contrasting them with competing views one can find in the ancient world. We then turn to the earliest use of symmetry in probability,

the decomposition into equipossible cases, as found in Bernoulli and Laplace, and the justifications for such a decomposition based on either logical criteria (the principle of indifference), or mathematical arguments (the method of arbitrary functions). We next take up the rule of succession (a consequence of a particular type of symmetry), which was the classical probabilistic justification for induction, and briefly trace its vicissitudes, from the ingenious arguments of Bayes to the pioneering efforts of the English logician W. E. Johnson. Johnson, before either de Finetti or Carnap, realized the utility of the concept of exchangeability (a more limited form of symmetry than equipossible cases) in an epistemic theory of probability. Johnson also recognized that rather than attempting to identify a unique probability function to describe belief, one could instead (and more modestly) use symmetry arguments to narrow down the field of possibilities to a family of functions (what Carnap later termed the continuum of inductive methods) described by a small number of parameters.

In recent decades such more circumscribed symmetry arguments have been extended in a number of very interesting ways as, for example, in the concept of partial exchangeability and the sampling of species paradigm; and these then are briefly discussed, as well as Jeffrey conditioning, a very different form of symmetry, being dynamic as opposed to static in nature (because the symmetry in question refers to epistemic states before and after a belief change).

There are, of course, many other ways in which symmetry arguments have entered into probabilistic reasoning. Some of these are illustrated by our final examples, a number of apparent paradoxes in both probability and inductive inference, all of which center around the sometimes casual and often siren-like appeal of symmetry. Symmetry arguments are a powerful tool in the arsenal of the mathematician, statistician, and philosopher, but they can often have unexpected consequences and caution is always appropriate.

15.1 The Taming of Chance

One of the great mysteries in the history of probability is the complete absence of a probability calculus in the ancient world. The use of randomization was widespread throughout all of its civilizations, for example in games of chance, divination, and jury selection, but there was no accompanying calculus of chance. This is certainly surprising in a civilization capable of Ptolemy's *Almagest*, and many theories have been advanced to explain the absence.

Could there have been conceptual impediments that explain this absence? The analysis of chance events based on a decomposition into equally likely cases is traditionally traced to the Pascal–Fermat correspondence of 1654, and is the basis of Abraham de Moivre's classic book *The Doctrine of Chances* (1718). Such an approach seems natural today, but in retrospect there may have been at least two subtle obstructions to its use in ancient times.

First, for the ancient Greeks and Romans, chance (Greek, *tyche*) or fortune (Latin, *fortuna*) meant the the unpredictable, and the unpredictable was thought to defy theorizing. Three quotations, one from each of three of the greatest ancient historians, will illustrate the point:

> So it is that chance always decides the greatest issues in human affairs in an arbitrary fashion. (Polybius, 2.70)

> But fortune, whose power is very great in all spheres, but particularly in warfare, often brings about great reversals by a slight tilt of the balance. (Caesar, *Civil War*, 3.68)

> I am undecided whether the affairs of human beings evolve by fate, and an immutable inevitability, or by chance [*forte*]. (Tacitus, *The Annals*, 6.22)

(Thus E. R. Dodds, 1951, p. 242, refers, with the Greeks in mind, to "the purely negative idea of the unexplained and unpredictable, which is Tyche".)

A second potential impediment to the use of modern symmetry arguments in the classical era was a common type of argument in Greek philosophy, viz.: *if the reasons in favor of two outcomes are equally balanced, neither should occur.* It was on some such grounds that Anaximander argued the Earth must lie at the center of the universe (Aristotle, *de Caelo* 295. b. 10; compare Plato, *Phaedo* 108e–109a and *Timaeus* 62d.12); and Parmenides that the universe must be eternal (inasmuch as there is no more reason for it to have been created at one time than another, *all* times are thereby ruled out); see Kirk and Raven (1957, p. 273), Owen (1966). A similar device is the use by the Greek Pyrrhonian skeptics of equally balanced arguments as a means of achieving *epoche*, or suspension of belief. The argument reappears later in Leibniz and others; see generally Zabell (1988, section 4).

Here symmetry arguments play the curious role of excluding possibilities rather than distributing probabilities. (Or, if you like, assigning zero probability to those possibilities.) Thus the modern use of symmetry in probability should not be regarded as inevitable, and marks a clear departure from the past.

15.2 Equipossibility

In the earliest theories of numerical probability, it is supposed there are a finite number of "chances", or equiprobable possibilities. The classic statement of this approach is given by de Moivre (1718/1756, pp. 1–2):

> [I]f we constitute a fraction whereof the numerator be the number of chances whereby an event may happen, and the denominator the number of all the chances whereby it may either happen or fail, that fraction will be a proper designation of the probability of happening.

Thus, setting aside the purely mathematical task of counting up the number of favorable and unfavorable "chances", the assignment of probability reduces to one of identifying the relevant chances. (De Morgan, 1845, pp. 395–6, particularly stresses this division of labor.)

15.2.1 James Bernoulli

In James Bernoulli's *Ars Conjectandi* of 1713 such chances are grounded in physical equipossibility: "I assume that all cases are equally possible, or can happen with equal ease" (1713/2006, p. 322). For example, in the case of a die, its faces

> all have equal tendencies to occur; because of the similarity of the faces and the uniform weight of the die, there is no reason why one of the faces should be more prone to fall than another,

and similarly for slips of paper drawn from an urn (p. 327). If this is not initially so, "then a correction must be made"; the cases must be further subdivided until equipossibility occurs:

> For any case that happens more easily than the others as many more cases must be counted as it more easily happens. For example, in place of a case three times as easy I count three cases each of which may happen as easily as the rest. [ibid]

This is clear enough, but illustrates the challenge of interpretation, because for Bernoulli probability is "a degree of certainty" (p. 315). In effect, Bernoulli's is an epistemic theory of probability grounded in physical equipossibility. The extension of this theory from games of chance to other areas such as medicine, meteorology, or the law, his ultimate goal, raises still further issues: just what are the assumed underlying equipossible cases here, and how can they be subject to calculation if they are hidden from us? The necessity of determining such hidden chances by a posteriori rather than a priori means led Bernoulli to formulate and prove the earliest version of the law of large numbers. (In this simple setting, this says that the proportion of successes in a independent sequence of trials performed under the same conditions converges to the true probability of success; see, e.g., Feller, 1968, pp. 152–3.) For further discussion of Bernoulli, see Hacking (1971b).

Such difficulties are already evident in John Arbuthnot's (1712) argument for Divine Providence on the basis of the excess of male over female christenings in London during an 82-year period: this could not be the result of "chance", Arbuthnot argued, since the probability of such a persistent excess occurring by chance is only $1/2^{82}$ (this is generally regarded as the first instance of a statistical test of significance). Critics (such as Nicholas Bernoulli) quickly retorted that Arbuthnot's conclusion did not follow: his data were consonant with chance, but chance arising from a die having 18 faces labelled *M* and 17 labelled *F*; see, e.g., Shoesmith (1985 and 1987).

But even with classical games of chance controversy was still possible—witness D'Alembert's delusion that the probability of getting at least one head in two tosses of a fair coin is 2/3 rather than 3/4. (D'Alembert argued there were three cases: both tails, both heads, and one each.) For an interesting classical discussion of D'Alembert, see De Morgan (1845, p. 401 and 1847, p. 173); for modern dissections, Daston (1979) and Swijtink (1986).

15.2.2 Laplace

Pierre Simon, the Marquis de Laplace, gave what may be regarded as the definitive statement of the epistemic theory of equipossible cases. Probability, for Laplace, is relative in part to our knowledge and in part to our ignorance, and this is reflected in how we measure it:

> The theory of chances consists of reducing all the events of the same type to a certain number of equally possible cases, that is to say, such that we are equally undecided about their existence. [Laplace, 1814, p. VIII]

Later on, at the beginning of Book II, Chapter 1, of the *Théorie analytique des probabilités* (1812/1886), Laplace adds the further gloss that we judge cases to be equipossible when we have no more reason to believe one will occur rather than another ("*lorsque rien ne porte à croire que l'un de ces cas doit arriver plutôt que les autres, ce que les rend, pour nous, également possibles*").

What if the cases are not initially "*également possibles*"? Then Laplace instructs us to divide and conquer: determine the respective possibilities of the cases ("*on déterminera leurs possibilités respectives*"), and use this to compute the probability of the original cases "relative to the subdivision of all the cases into the others that are equally possible". The probability of an event is then the sum of the probabilities of the favorable cases thus computed.

No general guidance is given regarding how to determine an appropriate decomposition into equally possible cases, only a caution: "The just appreciation of these different cases is one of the most delicate points in the analysis of chance"! For further discussion of early equipossibility theories of probability, see Hacking (1971a).

15.2.3 The Principle of Indifference

For Laplace cases are equipossible when we have no more reason to believe one than the other. This was a matter of judgment, even if at times a delicate one, and no general principle was stated. Indeed, a precise formulation defied general statement. Half a century later, Boole (1862, p. 229) referred ruefully to "that principle, more easily conceived than explained, which has been differently expressed as the 'principle of non-sufficient reason', the 'principle of the equal distribution of knowledge or ignorance', and the 'principle of order'." Keynes (1921/1973, p. 44) thought such terms infelicitous, and suggested the 'principle of indifference' instead.

The history of the vicissitudes of the principle of indifference throughout this period is a long and melancholy one. Ellis (1843) rejected epistemic probability outright, proclaiming *ex nihilo, nihil* (out of nothing, nothing). Others, like Boole, gave the principle qualified support in appropriate circumstances (see Zabell, 1989a, pp. 248–9). Johannes von Kries (1886, p. 11), one of the more incisive critics, argued that "the selection of the equally likely cases must be done in a compelling way and not be arbitrary" ("*die Aufstellung der gleich möglichen Fälle muss eine in zwingender Weise und ohne jede Willkür sich ergebende sein*"); and advanced a theory of *Spielräume* which he hoped would accomplish precisely this in particularly favorable situations. Keynes's *Treatise* (1921/1973, chapter 4) gives a good picture of this tortuous past history.

From a modern subjectivist standpoint, all this was a tempest in a teapot. But until a detailed operational account of epistemic probability in terms of betting odds or consistent beliefs was formulated (by Ramsey and de Finetti), so that the principle of indifference could be dispensed with as a source of the raw material of input probabilities, no resolution could be forthcoming (see Zabell, 1992). Indeed, Ramsey himself (1926/1964, pp. 85–6) saw the exorcism of the principle as one of the principal merits of his theory:

> [T]he Principle of Indifference can now be altogether dispensed with; we do not regard it as belonging to formal logic to say what should be a man's expectation of drawing a white or a black ball from an urn; his original expectations may within the limits of consistency be any he likes; all we have to point out is that if he has certain expectations he is bound in

> consistency to have certain others. This is simply bringing probability into line with ordinary formal logic, which does not criticize premisses but merely declares that certain conclusions are the only ones consistent with them. To be able to turn the Principle of Indifference out of formal logic is a great advantage; for it is fairly clearly impossible to lay down purely logical conditions for its validity, as is attempted by Mr Keynes. I do not want to discuss this question in detail, because it leads to hair-splitting and arbitrary distinctions which could be discussed for ever.

15.2.4 The Method of Arbitrary Functions

One of the most common arguments in favor of the physical interpretation of probability based on symmetry arguments is that coins really do seem to come up heads half the time, a roulette wheel really does come up red approximately 18 in 38 times, and so on. As Poincaré (1902/1905, pp. 213–14) noted,

> A gambler wants to try a *coup*; he asks my advice. If I give it to him, I shall use the calculus of probabilities, but I shall not guarantee success. This is what I shall call *subjective probability*. … But suppose that an observer is present at the game, that he notes all its *coups*, and that the game goes on a long time. When he makes a summary of his book, he will find that events have taken place in conformity with the laws of the calculus of probabilities. This is what I shall call *objective probability*, and it is this phenomenon which has to be explained.
>
> There are numerous insurance companies which apply the rules of the calculus of probabilities, and they distribute to their shareholders dividends whose objective reality can not be contested. To invoke our ignorance and the necessity to act does not suffice to explain them.
>
> Thus absolute skepticism is not admissible. We may distrust, but we can not condemn *en bloc*. Discussion is necessary.

Poincaré's solution to this conundrum was his *method of arbitrary functions*. Consider, for example, a wheel of fortune divided into successive red and black sectors. Apply an angular momentum $\omega > 0$ to the wheel: it spins and then comes to rest on either red or black. A little thought should make it clear that the ω-axis can be divided into successive intervals $J_1, J_2, J_3, \ldots$ such that throughout each interval the wheel always stops at the same color. Thus if the wheel starts in a red sector, then (letting R denote "red", B "black"),

$$P(R) = P(J_1) + P(J_3) + P(J_5) + \ldots,$$
$$P(B) = P(J_2) + P(J_4) + P(J_6) + \ldots.$$

Suppose $f(\omega)$ is a continuous probability density for ω. (That is, $f(\omega)$ is a nonnegative function whose total integral is one; the probability of an interval is then found by computing the integral of $f(\omega)$ over that interval. This probability can be interpreted as

being either physical or epistemic.) Then

$$P(R) = \sum_{k=0}^{\infty} \int_{J_{2k+1}} f(\omega)\, d\omega,$$

$$P(B) = \sum_{k=1}^{\infty} \int_{J_{2k}} f(\omega)\, d\omega.$$

Let $|J_k|$ denote the width of $|J_k|$. Poincaré demonstrated that in the limit as $\max |J_k| \to 0$, provided only successive $|J_k|$ are comparable in size and the function $f(\omega)$ is sufficiently smooth, $P(R)$ and $P(B) \to 1/2$. That is, for J_k small, the macroscopic probabilities $P(R)$ and $P(B)$ will be roughly equal, independent of the microscopic selection probability $f(\omega)$ for ω (the "arbitrary function"). The mathematical assumptions have natural interpretations: the J_k must be small enough to ensure we are unable to select any particular one; that successive J_k have comparable widths reflects the underlying physics; the smoothness of $f(\omega)$ reflects that small changes in ω should not result in large changes in the probability of ω.

Qualitatively, such arguments go back to von Kries (1886) and his "*Stoss-spiel*". Poincaré (1896) gave the first careful mathematical treatment; later Eberhard Hopf considerably added to the mathematical theory in the 1930s. The monograph by Eduardo Engel (1992) discusses and extends this earlier work. For historical discussions see von Plato (1983) and Heidelberger (2001). Recent years have seen renewed philosophical interest in the von Kries approach; see, e.g., Rosenthal (2010), Strevens (2011).

15.3 The Rules of Succession

Late in the 18th century the calculus of probabilities was extended to answer Hume's problem of induction: what is our logical ground or justification for thinking the future will resemble the past?. (See Hacking, 1975, chapter 19, for an interesting and provocative discussion of why Hume's problem arose when it did.) To understand the issues, first some terminology. Let $X_1, X_2, \ldots X_n$ be a sequence of 1s and 0s, recording respectively the occurrence or nonoccurrence of an event in a sequence of trials, so that $S_n = X_1 + X_2 + \cdots + X_n$ records the number of successes (occurrences of the event) in the n trials. If the outcomes are independent and have probability p of success, then the number of successes is governed by the *binomial distribution*:

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad 0 \le k \le n.$$

Here

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

is the *binomial coefficient*, and $n! = n \cdot (n-1) \cdot (n-2) \ldots 3 \cdot 2 \cdot 1$.

Suppose p is unknown, and $d\mu(p)$ is some probability distribution over the possible values of p. (Here $d\mu(p)$, commonly termed the "prior" or "initial" probability of p, can be either aleatory, the result of a physical process—say selecting a ducat from a bag of ducats, each having a different probability p of heads—or epistemic, a degree of belief about the different possible values of p. The mathematics is the same in either case.) Then one has

$$P(S_n = k) = \int_0^1 \binom{n}{k} p^k (1-p)^{n-k} d\mu(p), \qquad 0 \le k \le n.$$

But what is an appropriate choice for the prior? From a physical point of view, it is what it is. The prior is a physical fact, one subject to empirical determination or testing. But if the prior is epistemic, it must summarize our beliefs regarding the value of p. One common choice, used by Bayes and Laplace, was the so-called *uniform prior*: $d\mu(p) = dp$. What was its justification?

15.3.1 Bayes's Postulate

Suppose we know nothing about the probability of an event. For example, in the words of the Reverend Thomas Bayes (1764):

> [C]oncerning such an event I have no reason to think that, in a certain number of trials, it should rather happen any one possible number of times than another.

This is a *discrete* principle of indifference:

$$P(S_n = k) = \frac{1}{(n+1)}, \quad 0 \le k \le n.$$

Suppose we assume this holds (that is, accurately captures our beliefs) for all $n \ge 1$. Remarkably, this identifies $d\mu$: since for all $n \ge 1$ and $0 \le k \le n$ one has

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} dp = \frac{1}{n+1} \quad \text{(by mathematics)}$$

and

$$\int_0^1 \binom{n}{k} p^k (1-p)^{n-k} d\mu(p) = P[S_n = k] = \frac{1}{n+1} \quad \text{(by assumption)}$$

it follows that $d\mu(p) = dp$ by the Hausdorff moment theorem. (Of course Bayes, who preceded Hausdorff by a century and a half, did not know this argument, but it is interesting his intuition in the matter admits of so simple a justification.) For modern discussions of Bayes's paper, see Edwards (1978), Stigler (1982 and 1986, pp. 122–31), Hald (1998, chapter 8), Dale (1999, chapters 2–4).

15.3.2 Laplace's Rule of Succession

Bayes adopted a discrete indifference principle as capturing or defining total ignorance concerning p; and used it to deduce the continuous one. Laplace, whose first papers on

probability date from the 1770s, starts instead directly from *dp*. Proceeding either way, if $P(A|B)$ denotes the conditional probability of an event A given another event B, one has

$$P(X_{n+1} = 1|S_n = k) = \frac{k+1}{n+2}.$$

This was seen by many as a probabilistic explication for inductive inference, at least in the simple case of enumerative induction. It seemed that Hume had been silenced. (Indeed Richard Price, who was responsible for the posthumous publication of Bayes's essay, added an appendix to Bayes's essay discussing its application to the problem of induction. This was unquestionably intended as an answer to Hume; see Gillies, 1987 and Zabell, 1997.)

Laplace's "rule of succession" (the actual terminology is due to Venn, 1888, p. 190) almost immediately generated considerable controversy in the 19th century: John Stuart Mill, Richard Leslie Ellis, George Boole, and John Venn in England, Antoine Augustin Cournot and Joseph Bertrand in France, and Jacob Friedrich Fries in Germany all advanced criticisms of the Laplacean edifice, some general (e.g., probabilities are observed frequencies, not degrees of belief), but others more specific, in particular questioning the precise justification for the use of *dp* as the prior. (Later in the century Johannes von Kries was a particularly able critic.)

Several defenses were then mounted by different proponents of the Laplacean *Weltanschauung*. One of these was *empirical* (as argued, for example, by the English economist and statistician F. Y. Edgeworth): the uniform prior is often a good approximation to actual experience. Another might be termed *adaptive* (Louis Bachlier and the actuary G. F. Hardy): one can expand the class of priors to include ones which better describe our knowledge (in particular, the class of beta priors). And yet another was *logical* (W. E. Johnson): one can appeal to a more limited form of postulate, less sweeping in what it assumes. For further discussion of this 19th-century debate, and references to the relevant literature of the time, see Zabell (1989b).

15.3.3 W. E. Johnson: the Logical Breakthrough

The history of this debate could easily be the subject of a book. But in retrospect the key logical breakthrough was due to a Cambridge logician and economic theorist, William Ernest Johnson (1858–1931). Johnson, whose students included Keynes and Ramsey, was the author of a three-volume treatise on *Logic*, published in the 1920s, in which he first attacked the problem of induction, later returning to it in work that was published only after his death in three papers that appeared in the journal *Mind* in 1932.

15.3.3.1 *The multinomial generalization*

Johnson contributed four key ideas to this protracted debate. The first was to start from the case of *multinomial* outcomes: instead of restricting himself to just two possible outcomes (the event does or does not occur), he considered the more general case of $t \geq 2$ possible outcomes. This was an important move, because although in some cases one might (justifiably) ridicule the idea that an event and its negation were equally likely, one might instead, as envisaged by Bernoulli, be in a situation (at least at some ultimate

underlying level) where there was indeed a set of equally likely alternatives. (Mathematically, the multinomial case had already been discussed by Laplace and later writers, but the point here is its use as part of a principled philosophical attack on the problem of induction.) If the t different outcomes are denoted $c_1, \ldots, c_t$, then the basic quantities of interest are the probabilities of the form

$$P(X_1 = e_1, X_2 = e_2, ..., X_n = e_n)$$

for $e_i \in \{c_1, ..., c_t\}, 1 \leq i \leq n$. (The displayed quantity denotes the probability that simultaneously $X_1 = e_1$, and $X_2 = e_2, \ldots$, and $X_n = e_n$.) The sequence $(e_1, \ldots, e_n)$ is what Carnap later termed a "state description".

15.3.3.2 *The permutation postulate (1924)*

The second element introduced by Johnson was the *permutation postulate* (1924): this assumes that for every possible sequence of outcomes $e_1, ..., e_n$ and every permutation σ of $\{1, ..., n\}$, one has

$$P(X_1 = e_1, ..., X_n = e_n) = P(X_1 = e_{\sigma(1)}, ..., X_n = e_{\sigma(n)}).$$

The permutation postulate is a symmetry assumption respecting the order of outcome. For example, it expresses the belief that in ten tosses of a coin, any sequence of six heads and four tails is as likely as any other. It is the same as Bruno de Finetti's concept of *exchangeability*, introduced independently by de Finetti (1931/1989) several years later. (An *exchangeable sequence* is therefore a sequence $X_1, \ldots, X_n$ which satisfies the permutation postulate.)

Important summary statistics here are the *frequencies* n_j: the number of e_i in state c_j. If the *multinomial coefficient* is defined to be

$$\binom{n}{n_1\ n_2\ \ldots\ n_t} = \frac{n!}{n_1!\, n_2! \ldots n_t!},$$

then a fundamental quantity of interest of interest is

$$P(n_1, n_2, ..., n_t) = \binom{n}{n_1\ n_2\ \ldots\ n_t} P(X_1 = e_1, ..., X_n = e_n).$$

The equality holds because every sequence $e_1, \ldots, e_n$ having the same frequency counts has the same probability, the multinomial coefficient records the number of such sequences for a given set of frequency counts, and the formula follows immediately from the additivity of probability. The vector of frequency counts $(n_1, \ldots, n_t)$ is what Carnap later termed a "structure description".

The finite representation theorem

The simplest example of a finite exchangeable sequence is sampling without replacement from a finite population having a known composition. Think of an urn $U_{n,r}$ containing r red balls and $n - r$ black. If the n balls are drawn one after another at random, there are $\binom{n}{r}$ ways this can happen (that is, possible sequences of r red balls and $n - r$ black balls), each equally likely to occur. Denote such a probability distribution $H_{n,r}$; it is, by definition, exchangeable.

More generally, suppose (p_r) is a probability distribution on $0 \leq r \leq n$. Then it is not hard to verify that the mixture $\sum_r p_r H_{n,r}$ is an exchangeable probability on n-long sequences of reds and blacks; it corresponds to first choosing an urn $U_{n,r}$ according to the distribution (p_r), and then sampling without replacement from it. (The verification of this assertion comes down to a simple application of the theorem of total probability: if an event can be decomposed into disjoint, mutually exclusive subevents, then the probability of the event is the sum of the probabilities of the constituent subevents.)

This is in fact the structure of the general finite exchangeable sequence P on sequences of length n. If S_n is the number of reds observed in the sequence, then given S_n one knows the number of reds in the sequence but not their order. Thus the conditional distribution of P given S_n is the $H_{n,r}$ distribution, hence (by a conditional form of the theorem of total probability) if $p_r = P(S_n = r)$, one has $P = \sum_r p_r H_{n,r}$.

The corresponding representation for multinomial sequences is just as simple. If $\mathbf{n} = (n_1, \ldots n_t)$ is a structure description, and $n = n_1 + \cdots + n_t$, then $H_{\mathbf{n}}$, the probability distribution for sampling without replacement from a known population containing n_j members of each type j, is exchangeable; and if P is an n-long exchangeable sequence such that $P(\mathbf{n}) = p_{\mathbf{n}}$, then $P = \sum_{\mathbf{n}} p_{\mathbf{n}} H_{\mathbf{n}}$.

This is the *de Finetti representation* for a finite exchangeable sequence. Previously one argued by analogy, comparing successive observations to random selection from the "urn of nature" (see generally Strong, 1976). For example, in a celebrated passage, William Stanley Jevons wrote:

> Nature is to us like an infinite ballot box, the contents of which are being continually drawn, ball after ball, and exhibited to us. Science is but the careful observation of the succession in which balls of various character present themselves [Jevons, 1877, p. 150].

Question the aptness of this analogy, and the entire edifice comes crashing down. The significance of the de Finetti representation for the theory of inductive inference is that *it is no longer necessary to refer to the urn of nature*. The importance of this cannot be overemphasized.

The use of exchangeability reverses this process of arguing by analogy with drawing from an urn of unknown composition. Suppose you are in a setting in which simple enumerative induction is thought appropriate (that is, one based on simply counting the number of favorable and unfavorable outcomes in a sequence of observations). Then (assuming you accept the existence of epistemic probabilities in the first place, more about this shortly) your probability assignment must be exchangeable. (For if not, then for at least one structure description $(n_1, \ldots, n_t)$ some sequences are more likely than others; and this implies the presence of information in addition to that usually posited, so that simple enumerative induction is no longer appropriate.) But if the sequence is exchangeable, then the representation theorem tells us that $P = \sum_{\mathbf{n}} p_{\mathbf{n}} H_{\mathbf{n}}$; that is, our probability assignment is *as if* we were drawing balls from an urn of unknown composition, and we assign probabilities $p_{\mathbf{n}}$ to the different possible (but unknown) compositions $\mathbf{n}$. But now the "as if" is a mathematical consequence of the assumption of exchangeability (which mirrors our present state of knowledge), and not some vague appeal to an "urn of nature".

Critics of the rule of succession had noted that some assumption was needed in order for a simple enumerative inductive inference to be valid (Mill's "uniformity of nature", Keynes's "principle of limited variety", and later, of course, Goodman's "projectability", see Goodman,

1946 and 1979, Stalker, 1994). It was Johnson's achievement to have realized that all such vague notions could be replaced by the mathematically precise concept of exchangeability.

15.3.3.3 *The combination postulate (1924)*

The third element in Johnson's initial analysis was his introduction of *the combination postulate*: this posits that *all frequency vectors* $(n_1, n_2, \ldots n_t)$ *have the same probability*. Granting this, it follows that

$$P(n_1, n_2, ..., n_t) = \frac{1}{\binom{n+t-1}{n}}.$$

(Feller, 1968, terms these "Bose-Einstein statistics", and discusses some of their properties.) Note this is the natural generalization of Bayes's postulate for the case $t = 2$, for then the structure description reduces to:

$$(n_1, n_2) = (k, n-k).$$

The combination postulate is another form of symmetry assumption, and a very strong one at that. Indeed, the combination (and permutation) postulates together imply that

$$P(X_1 = e_1, ..., X_n = e_n) = \frac{1}{\binom{n+t-1}{n}\binom{n}{n_1\ n_2\ \ldots\ n_t}}$$

(this is Carnap's $m^\star$ function); and also that

$$P(X_{n+1} = c_i \mid n_1, n_2, ..., n_t) = \frac{n_i + 1}{n+t}$$

(this is Carnap's $c^\star$ function). See Carnap (1950) for $m^\star$ and $c^\star$. If all this were immune to question, then Hume would indeed be vanquished and unique inductive probabilities would exist. But criticisms of the combination postulate were soon raised, and Johnson came to have reservations about the combination postulate (just as Carnap came to have as well, even before the publication of his *Logical Foundations of Probability* in 1950). For one contemporary critic, see Broad (1924).

15.3.3.4 *The sufficientness postulate (1932)*

Suppose for any $n \geq 1$, the sequence $X_1, X_2, \ldots, X_n, \ldots$ is exchangeable (that is, $X_1, \ldots, X_n$ is exchangeable for every $n \geq 1$) and also satisfies the following three conditions:

1. Any state description $e, ..., e_n$ is possible: $P(X_1 = e_1, ..., X_n = e_n) > 0$.
2. The "sufficientness postulate" is satisfied: namely, after n observations, the probability of seeing the next outcome fall into the i-th category depends only on how many previous outcomes fell into that category and how many did not. That is, for some

function f (that depends only on n_i and n, and not on the individual outcomes) one has

$$P(X_{n+1} = c_i \mid n_1, ..., n_t) = f(n_i, n).$$

3. There are at least three types of species: $t \geq 3$.

Then Johnson was able to prove (1932, see Zabell, 1982) the following beautiful result:

Theorem 1: Under the above three conditions on the exchangeable sequence $X_1, X_2, \ldots$, either (1) the sequence is independent, or (2) there exists a positive constant α such that

$$P[X_{n+1} = c_i \mid n_1, ..., n_t] = \frac{n_i + \alpha}{n + t\alpha}$$

for every $n \geq 0$, and set of frequency counts $n_1, ..., n_t$.

This is Carnap's (1952) "continuum of inductive methods". (For discussion of Carnap's continuum, see Zabell, 2007 and 2011.)

Johnson's "sufficientness postulate" (the terminology is due to I. J. Good) is yet another form of symmetry assumption: the conditional probability of seeing an outcome of the j-th type on the next trial depends only on the total number of trials n and the number of outcomes n_j of the type seen. The postulate seems relatively innocuous as a way of capturing ignorance regarding categories but (as Alan Turing realized during World War II, in conjunction with his cryptological attack on the German naval Enigma machine), there do exist situations in which we know nothing about the individual types, and yet the frequencies $n_1, \ldots, n_t$ do contain information for predicting the j-type, beyond just n_j itself; see Good (1965, p. 6), Zabell (1996).

Beautiful as Johnson's results are, do not his assumptions suffer from the same charge of arbitrarines as was leveled against the principle of indifference? The answer is *no*, given the spirit in which they were proposed. Johnson tells us:

> the postulate adopted in a controversial kind of theorem cannot be generalized to cover all sorts of working problems; so it is the logician's business, having once formulated a specific postulate, to indicate very carefully the factual and epistemic conditions under which it has practical value. [Johnson, 1932, pp. 418–19]

In short, there are no universally applicable postulates: different symmetry assumptions are appropriate under different circumstances, and none is logically compulsory. The best one can do is identify symmetry assumptions that seem natural, have identifiable consequences, and may be a natural reflection of one's beliefs under some reasonable set of circumstances. In judging the appropriate use of the sufficientness postulate, for example, the issue is whether or not you think the postulate accurately captures the epistemic situation at hand.

Remark

There is a beautiful extension of the Johnson-Carnap continuum in which predictive probabilities are permitted to depend on T, the number of categories observed thus far;

see Hintikka (1966), Hintikka and I. Niiniluoto (1980). Kuipers (1978) is an outstanding summary of a variety of such continua and their interconnections.

15.4 The de Finetti Representation Theorem

This raises the question of what one can say under the weaker assumption of just exchangeability. Let us say that an infinite sequence $X_1, X_2, \ldots$ is *infinitely exchangeable* if the finite sequence $X_1, \ldots, X_n$ is exchangeable for every $n \geq 1$. In the 1930s the Italian mathematician Bruno de Finetti was able to prove (see, e.g., de Finetti, 1937/1964) the following important result:

Theorem 2. *Let $X_1, X_2, \ldots$ be an infinitely exchangeable sequence of 0s and 1s, and let $S_n := X_1 + \ldots + X_n$. Then*

1. *The limit $Z := \lim_{n\to\infty} \frac{S_n}{n}$ exists with probability one;*
2. *If μ is the probability distribution of the limiting frequency Z (so that for every event A, $\mu(A) = P(Z \in A)$), then for every $n \geq 1$ and $0 \leq k \leq n$,*

$$P(S_n = k) = \binom{n}{k} \int_0^1 p^k (1-p)^{n-k} d\mu(p).$$

De Finetti's beautiful result simultaneously provides:

(1) *A subjective explanation for the existence of objective chance*. The existence of limiting frequencies (the Z) emerges as a mathematical consequence of the qualitative symmetry assumption of exchangeability, rather than as a dubious (in part because untestable) physical assumption about the existence of infinite limiting frequencies.

(2) *A rationale for using priors in statistical inference* (they correspond to our degrees of belief regarding the distribution of the limit Z). The representation theorem tells us that in the exchangeable setting an agent is acting as though spreading credences over various particular chance hypotheses, their weights given by $d\mu$, hypotheses according to which the trials are independent, with constant chances.

(3) *A partial solution to Hume's problem of induction*. This is because in this setting most posterior distributions will concentrate about a single value of p as n increases, and so predicted future frequencies will mirror those of the past. (The "most" here is intended to rule out what one might term dogmatic priors in which stretches of the unit interval $0 \leq p \leq 1$ are ruled out ahead of time as being impossible, by assigning them probability zero. An entire paper could be written on this issue, and it would take us too far afield to go into it further here.)

Caution: De Finetti was a finitist and had more nuanced views; see Zabell (2011) for further discussion of this point.

15.4.1 Partial Exchangeability

An *infinitely exchangeable sequence* $X_1, X_2, X_3, \ldots$ (ad infinitum), is one in which for each $n \geq 1$, the probability distribution of the initial segment $X_1, \ldots, X_n$ is invariant under all permutations of the time index $1, \ldots, n$. De Finetti's theorem admits of a sweeping generalization in this setting. Under only mild restrictions on the outcome space (rather than the simplest case of 0s and 1s) *every* infinitely exchangeable sequence is necessarily a mixture of independent and identically distributed random variables. ("Random variables" here means variables for which more general values are taken rather than just 0 and 1. Aldous, 1985, provides an excellent overview of such results.)

This suggests a possible program: to find additional conditions leading to a mixture of one of the parametric families of classical statistics. Consider, for example, the celebrated "bell-shaped" normal (or Gaussian) distribution having density function

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right],$$

(where μ and $\sigma^2 > 0$ are two parameters specifying the mean and variance of the distribution, respectively.) Then one could ask: when is an exchangeable sequence a mixture of normally distributed random variables? More generally, can one find alternative forms of exchangeability appropriate to, say, mixtures of Markov chains (where successive outcomes are not independent)? The term "partial exchangeability" (de Finetti, 1938/1980) refers to such conditions; see Diaconis and Freedman (1980). For extensions of Carnap's program, see Skyrms (1991 and 1993); for a Johnson-type theorem for Markov chains, see Zabell (1995).

Recent decades have seen a beautiful circle of ideas dramatically extending such notions. For example, there is a simple and attractive characterization of mixtures of normally distributed random variables centered at 0 (have $\mu = 0$). The statement of this characterization requires the introduction of a little preliminary terminology. A linear transformation $U : \mathbf{R}^n \to \mathbf{R}^n$ is said to be *orthogonal* if it preserves the length of every vector in $\mathbf{R}^n$. (The reason for the term "orthogonal" is that this in turn implies that angles between pairs of vectors are also preserved, and in particular, the relation of being mutually perpendicular. Such transformations are the n-dimensional generalization of the rotations in three dimensions.) The sequence of random variables $X_1, \ldots, X_n$ is said to be *orthogonally invariant* if the random vector $U(X_1, \ldots, X_n)$ (which by definition of U, takes values in $\mathbf{R}^n$) has the same distribution for every orthogonal linear transformation U. (That is, the distribution of $X_1, \ldots, X_n$ is invariant under n-dimensional rotations.) For example, any permutation of $X_1, \ldots, X_n$ is an orthogonal transformation (of a very special kind); thus orthogonal invariance entails exchangeability but is much more restrictive. An infinite sequence $X_1, X_2, X_3, \ldots$ is then said to be orthogonally invariant if for every $n \geq 1$ the finite

initial sequence $X_1, \ldots, X_n$ is orthogonally invariant. In such cases one has the following very interesting result.

Theorem 3. *Every orthogonally invariant infinite sequence of random variables $X_1, X_2, X_3, \ldots$ is a mixture of independent and identically distributed sequences of normally distributed random variables with mean 0 and some variance σ^2.*

Let P_σ denote the distribution of an infinite sequence of independent and identically distributed random variables each having a normal distribution with parameters $\mu = 0$ and $\sigma^2 > 0$. Then this result tells us that there exists a probability measure Q (playing the same role as the mixing measure in the de Finetti representation for 0, 1 valued sequences) on the positive real numbers $\sigma > 0$ such that the distribution P for the infinite orthogonally invariant sequence takes the form

$$P = \int_0^\infty P_\sigma \, dQ(\sigma).$$

In the original de Finetti representation theorem, you chose p according to some distribution μ, and then tossed a "p-coin" an infinite number of times, and the result was the arbitrary exchangeable 0, 1 sequence. Here, you choose σ according to the probability distribution Q, and then generate a sequence of independent and identically distributed random variables $X_1, X_2, X_3, \ldots$, each having a normal distribution with parameters $\mu = 0$ and $\sigma^2 > 0$; the result is the general orthogonally invariant exchangeable sequence.

Just as the sample frequencies n_j play an important role in the analysis of the multinomial case, here too certain "sufficient statistics" play an important role. (In mathematical statistics, a statistic is sufficient if, informally, it contains the whole of the relevant sample information for purposes of estimating parameters, and, formally, is a statistic having the property that conditional on it, the resulting conditional distribution is independent of the parameters.) Here the relevant statistic is:

$$T_n = \sqrt{X_1^2 + \ldots + X_n^2}.$$

It can then be shown that the property of orthogonal invariance is equivalent to:

> for each $n \geq 1$, conditional on T_n the distribution of $X_1, .., X_n$ is uniform on the $n-1$-sphere of radius T_n.

Furthermore,

1. the limit $T = \lim_{n\to\infty} T_n/\sqrt{n}$ exists with probability one;
2. $P(T \leq \sigma) = Q([-\infty, \sigma))$; the measure Q is the limiting distribution of T.

In short, the classical parameter σ emerges as the almost sure limit of $T_n/\sqrt{n}$, the estimated standard deviation, and assigning a prior probability to σ comes down to quantifying our degree of belief concerning this large-sample limiting value. This suggests a more general program for Bayesian statistics: to explain the classical parametric families of mathematical statistics in terms of similar symmetry assumptions on the sequence of observations, the parameters emerging as the limiting values of appropriate sufficient statistics, and the priors of applied Bayesian statistics emerging as the degrees of belief regarding these limiting values. Diaconis and Freedman (1985) carries out this program in a number of important cases.

15.5 The Sampling of Species Problem

Suppose we do not know beforehand the identity of the possible species that can arise, but can only distinguish between them after the fact. ("This is the same as that; but this is different from that.") This is sometimes termed the *sampling of species problem*.

Consider the following three axioms, the first two of which parallel and the last replaces those of Johnson.

1. All sequences are possible (have positive probability).
2. The probability of seeing the i-th species already seen on the next trial is a function of the number of times that species has been observed, n_i, and the total sample size n: $f(n_i, n)$.
3. The probability of observing a new species depends only on the number of species already observed t, and sample size n: $g(t, n)$.

In this case it is possible to derive a *continuum of inductive methods for the sampling of species*. (In the following α and θ refer to two parameters analogous to those in the Carnap continua.)

Case 1 (more than one species is observed): If $n_i < n$ for some i, then there exist α and θ (independent of n) such that

$$f(n_i, n) = \frac{n_i - \alpha}{n + \theta}, \qquad g(t, n) = \frac{t\alpha + \theta}{n + \theta}.$$

There is a simple heuristic explanation for these formulas. If t is large, that is, a large number of species have already been observed, then by a form of enumerative induction, it is natural to expect to see yet more species in the future; and indeed, the predictive probability that this is the case, $g(t, n)$, is a linear function of t. But in order for the probabilities to add to one, one must subtract α from each of the remaining predictive probabilities $f(n_i, n)$ for the species. (Note that if $n_i < n$, then $t > 1$, so that there are at least two species, and the universal generalization is *dis*confirmed.) Not all parameter values are possible: it follows from the derivation that one must have (γ being a parameter appearing in Case 2)

$$0 \leq \alpha < 1; \quad \theta > -\alpha; \quad 0 \leq \gamma < \alpha + \theta.$$

Case 2 (only one species is observed): If $n_i = n$ for some i, then

$$f(n_i, n) = \frac{n_i - \alpha}{n + \theta} + c_n(\gamma), \qquad g(t, n) = \frac{t\alpha + \theta}{n + \theta} - c_n(\gamma);$$

here $c_n(\gamma)$ is the increase in the probability of seeing the i-th species again due to the confirmation of the universal generalization. The quantity $c_n(\gamma)$ is given by the following somewhat intimidating formula (which of course requires no small amount of mathematical

derivation):

$$c_n(\gamma) = \frac{\gamma(\alpha+\theta)}{(n+\theta)\left[\gamma + (\alpha+\theta-\gamma)\prod_{j=1}^{n-1}\left(\frac{j-\alpha}{j+\theta}\right)\right]}.$$

The important thing here is not so much the specific form of the formula, however, but the fact that it exists! That is, that the three postulates given above completely determine the form of the predictive probabilities, up to the two parameters α, θ, and γ.

Once again, there is a simple heuristic explanation for the underlying form of the predictive probabilities. Suppose all n outcomes to date fall into a single category. Then there are two possibilities: 1) other categories can (and will eventually) occur, but this particular category likely has a high probability of occurring (which is why it has been seen so much); or 2) this is the *only* category that will occur (this is the universal generalization part). The observations to date support both of these possibilities, and both contribute to the predictive probability of the next observation falling into the category. But by how much? This is precisely what is given by the sum $f(n_i, n) = \frac{n_i - \alpha}{n+\theta} + c_n(\gamma)$.

For discussion of this continuum and, more generally, the analysis of the sampling of species problem in terms of Kingman's theory of *exchangeable random partitions*, see Zabell (1992 and 1997).

15.6 Jeffrey Conditioning

Richard Jeffrey's *probability kinematics* (Jeffrey, 1983, chapter 11) focuses on yet another form of symmetry, in this case one over time. Suppose a set of events $E_1, E_2, \ldots, E_n$ form a partition of the space of outcomes (that is, are mutually exclusive and exhaustive). Let P be an initial probability of an agent; P^* a later probability reflecting some change in the beliefs of the agent (it is *not* assumed these arise via classical Bayesian conditioning).

Jeffrey posits the partition $\{E_1, E_2, \ldots E_n\}$ of the sample space captures the totality of our belief change, in the sense that

$$P(E_i) \to P^*(E_i), \qquad 1 \le i \le n;$$
$$P(A \mid E_i) = P^*(A \mid E_i), \qquad 1 \le i \le n, \text{and all } A.$$

Thus, any change is possible on the E_i, but conditional probabilities relative to them do not change. In that case it then follows from the theorem of total probability that for any event A,

$$P^*(A) = \sum_{j=1}^{n} P^*(A|E_j)P^*(E_j)$$
$$= \sum_{j=1}^{n} P(A|E_j)P^*(E_j).$$

That is—under the assumed conditions—the final probability P^* is *entirely determined* by the values it assigns to the partition.

Jeffrey's probability kinematics is intended as a model of belief revision in non-Bayesian settings where the belief change cannot be realistically viewed as arising from learning with certainty a proposition or event. Despite its simplicity, it has a surprisingly rich set of mathematical consequences; see, e.g., Diaconis and Zabell (1982). Jeffrey (1988) provides a readable overview. For a discussion of probability kinematics as an instance of symmetry arguments in probability, see van Fraassen (1989, chapter 13).

15.7 A Budget of Paradoxes

Probability theory is enriched by a set of apparent "paradoxes" illustrating the complexities that can arise in applying the theory. Several of these involve the use or misuse of seemingly innocent symmetry assumptions.

15.7.1 The Three-Prisoner Paradox

Suppose that A, B, and C are three prisoners who have been sentenced to death, and that one has been pardoned (but it is not known who).

Prisoner A tells the warden: "Since there is only one pardon, I know at least one of B and C will be executed. Therefore you will not really be giving me any information if you tell me the name of one of them who is to be executed." The warden accepts this and says, "O. K. One of the two who is to be executed is B."

Now prisoner A beams. Before his chances of a pardon were 1 in 3. But now, with B eliminated, his chances have increased to 1 in 2!

The resolution of this paradox depends on the correct choice of sample space. (Put another way, Prisoner A should condition on his *total* evidence, and the sample space needs to be rich enough to reflect this.) It is often thought that there are three equally likely cases, depending on whether A, B, or C receives the pardon. But this overlooks what can happen if A has been pardoned: unlike the cases when B or C is pardoned (in which case the warden is constrained to saying *only* C or B, respectively); in the case of A being pardoned, the warden has the option of saying *either* B or C. Suppose in this last case the warden says "B" with probability p and "C" with probability $1-p$. Then a straightforward application of Bayes's theorem (see Stigler, 1986, pp. 102–5 for a statement and historical context) tells us that the probability of A being pardoned, given that the warden says "B", is

$$\frac{\frac{1}{3}p}{\frac{1}{3}+\frac{1}{3}p} = \frac{p}{1+p}.$$

Thus if $p = 1/2$ (whenever the warden has a choice, he chooses "B" or "C" with equal probability), the probability that A is pardoned indeed remains unchanged at $1/3$. But note this is the *only* value of p for which this is the case; for every other value, the warden's

statement conveys information to A and affects the latter's probability of being pardoned. In particular, when $p = 0$ (if the warden has a choice, he always says "C"), the probability drops to 0; and when $p = 1$ (if the warden has a choice, he always says "B"), the probability increases to $1/2$. (Note incidentally that even in the $p = 1/2$ case, although A's probability remains unchanged, C's has increased to $2/3$.)

The three-prisoner paradox is equivalent to the earlier (but less piquant) *Bertrand box paradox* (Bertrand, 1889): there are three boxes, one containing two gold coins, one containing two silver coins, and one containing one gold coin and one silver coin. One of the three boxes is chosen at random and a coin removed from it (also at random). If the coin selected is gold, what is the probability the remaining coin in the box is also gold? The correct answer is $2/3$, but a naive appeal to symmetry is often used to argue in favor of $1/2$. Thus stated, the problem seems straightforward enough, but it is logically equivalent to both the three-prisoner paradox and the *Monty Hall problem*, and in these later forms has generated considerable debate over the years; see Falk (1992), Rosenhouse (2009). Important elements in resolving such paradoxes include the proper identification of the sample space, and appropriate modeling of the receipt of information.

15.7.2 Bertrand's Paradox

Given a circle, what is the probability that the length of a random chord of the circle exceeds the side of the inscribed equilateral triangle? Solving this requires a precise definition of what "random chord" means. Bertrand provided three definitions, each giving a different answer! The three approaches are

- Choose a random point inside the circle, use it as the midpoint.
- Choose a random pair of points on the circle, use them as endpoints.
- Choose a random radius, then a random point on it, and use it as midpoint.

("Random" here means the obvious uniform probability distribution over the operative set.)

The resulting answers are $1/4$, $1/3$, and $1/2$, respectively. The point is choosing a chord "at random" depends on how chords are parametrized. Nevertheless, the physicist Ed Jaynes has argued (2003, pp. 386–96) that the problem does indeed have a unique solution *given its statement*, by appealing to symmetry considerations. Jaynes argues as follows: let R be the radius of the circle, (r,θ) the midpoint of the chord, and $f(r,\theta)$, the density function of the midpoint.

Given the statement of the problem, three kinds of symmetry are implicit: the density $f(r,\theta)$ should satisfy:

- rotational invariance;
- scale invariance;
- translational invariance.

Jaynes then shows purely mathematically that these forms of invariance together imply that

$$f(r,\theta) = \frac{1}{2\pi Rr}, \qquad 0 \le r \le R, \quad 0 \le \theta \le 2\pi\,.$$

(Boole would no doubt have found this very satisfying.) This corresponds to Bertrand's third definition of "at random", and gives a probability of 1/2. For an exposition and critique of Jaynes's argument, and its connection with his use of transformation groups as part of his maximum entropy approach to "objective Bayesian" inference, see Howson and Urbach, 2006, pp. 280–5.

This attractive example illustrates a much more general phenomenon: the use of group invariance to restrict the range of possibilities plays a major role in modern statistics. It is, for example, central to the approach of Sir Harold Jeffreys (1939); is the basis of Donald Fraser's "structural inference" (Fraser, 1968); and enters in an essential way in the theory of invariant and equivariant estimators (see, e.g., Lehmann and Cassela, 1998).

Questions about the definition of "random" in the continuous setting were at the heart of a number "design" controversies, as in the Forbes-Michell dispute (see Gower, 1982).

15.7.3 The Exchange Paradox

The following paradox has also been the subject of considerable discussion. Suppose A, B, and C play the following game.

1. C puts x euros in one envelope, $2x$ in another envelope.
2. One of the two envelopes is handed to A, the other to B.
3. A opens his envelope and sees that there are y euros in it.

Should A exchange envelopes if he can? Here is an argument that A should say yes. Let $|B|$ be the amount in B's envelope. A reasons:

$$P(|B| < y) = P(|B| > y) = \frac{1}{2},$$

so if I exchange I expect

$$\frac{y}{2} \cdot \frac{1}{2} + (2y) \cdot \frac{1}{2} = \frac{5y}{4} > y;$$

thus A offers to exchange.

The paradox: B also accepts the offer to exchange, since he reasons in similar vein; and yet surely it cannot be advantageous for both to exchange.

Dozens of papers have been written on this paradox over the last three decades; see, e.g., Katz and Olin (2007). Let us mention just one surprising twist here. Is it possible for a Bayesian to believe that conditional on the event $|A| = y$, it is equally possible for $|B|$ to be either less than or greater than $|A|$ (and this for every value of y)? That is, that

$$P(|B| < y \mid |A| = y) = P(|B| > y \mid |A| = y) = \frac{1}{2}$$

for all $y > 0$? The answer is *yes*, but only if you allow the use of genuinely finitely additive probabilities! (That is, probabilities which are only finitely additive, but *not* countably

additive.) Bruno de Finetti would have had no problem with this, of course, but many (most?) subjectivists today would be loath to part with the useful assumption of countable additivity. For two particularly lucid expositions of de Finetti's philosophy and approach, see de Finetti (1931/1989 and 1937/1964).

15.7.3.1 *Bruce Hill's A_n*

This last observation is closely related to an ignorance postulate for finite sampling due to the Bayesian statistician Bruce Hill. Let $X_1, X_2, X_3, \ldots$, be an exchangeable sequence. If the X_j are continuous (so ties have only a zero probability of occurring), then the first n X_j will divide the real line into $n+1$ disjoint subintervals. Hill's postulate is:

$$A_n: \quad X_{n+1} \text{ is equally likely to lie between any of the } X_j, 1 \leq j \leq n.$$

If $n = 1$, this says that the two events $X_2 < X_1$ and $X_2 > X_1$ are equally likely; that is, have probability 1/2. This is, of course, exactly the intuition underlying the exchange paradox. In his original paper Hill (1968) showed that if one restricts oneself to countably additive probabilities, no exchangeable sequence $X_1, X_2, X_3 \ldots$ can satisfy A_n; Lane and Sudderth (1978), on the other hand, were later able to show that finitely additive probabilities satisfying A_n exist for all n.

15.8 Conclusion

Since its inception symmetry arguments have played an integral role in the theory of probability. A "just appreciation" of the chances (as Laplace said) and a decomposition into equipossible cases appeared an adequate foundation for that theory as long as its realm of application remained limited. But the extension of the theory to law, medicine, and other practical endeavors called the appropriateness of this paradigm into question, and the assumptions of symmetry eventually became more modest. Today symmetry arguments remain valuable tools in narrowing the range of possible probability functions, but no more, and are recognized as such.

References

Aldous, D. (1985) Exchangeability and related topics. In Hennequin, P. L. (ed.) *École d'Été de Probabilités de Saint-Flour XIII*. Lecture Notes in Mathematics. Vol. 1117. pp. 1–198. New York, NY: Springer-Verlag.

Arbuthnot, J. (1712) An argument for divine providence, taken from the constant regularity observed in the births of both sexes. *Philosophical Transactions of the Royal Society of London*. 27. pp. 189–90.

Bayes, T. (1764) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*. 53. pp. 370–418.

Bernoulli, Jakob (1713/2006) *The Art of Conjecture*. Translated from the Latin by E. D. Sylla. Baltimore, MD: Johns Hopkins Press. (Originally published in Basel.)

Bertrand, J. (1889) *Calcul des probabilités*. Paris: Gauthier-Villars.

Boole, G. (1862) On the theory of probabilities. *Philosophical Transactions of the Royal Society of London*. 152. pp. 225–52.

Broad, C. D. (1924) Mr. Johnson on the logical foundations of science. *Mind*. 33. pp. 242–61, 369–84.

Carnap, Rudolph (1950) *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press. (2nd edition 1962.)

Carnap, Rudolph (1952) *The Continuum of Inductive Methods*. Chicago, IL: University of Chicago Press.

Cournot, Antoine Augustin (1843) *Exposition de la théorie des chances et des probabilités*. Paris: Libraire de L. Hachette.

Dale, Andrew I. (1999) *A History of Inverse Probability: From Thomas Bayes to Karl Pearson*. 2nd edition. New York, NY: Springer.

Daston, Lorraine J. (1979) D'Alembert's critique of probability theory. *Historia Mathematica*. 6. pp. 259–79.

de Finetti, Bruno (1931/1989) Probabilismo. *Logos*. 14. pp.163–219. (English translation Di Maio, Maria Concetta, Galavotti, Maria Carla, and Jeffrey, Richard C., *Erkenntnis*. 31. pp. 169–223.)

de Finetti, Bruno (1937/1964) La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*. 7. pp. 1 68. (Translated as "Foresight: its Logical Laws, its Subjective Sources" in Kyburg, H. E., Jr. and Smokler, H. E. (eds.) *Studies in Subjective Probability*. pp. 93–158. New York, NY: Wiley).

de Finetti, Bruno (1938/1980) Sur la condition de "équivalence partielle". In *Actualités Scientifiques et Industrielles*. Vol. 739. pp. 5–18. Paris: Hermann. (Translated by P. Benacerraf and R. Jeffrey as "On the condition of partial exchangeability", in R. C. Jeffrey (ed.) *Studies in Inductive Logic and Probability*. Vol. 2. pp. 193–205. Berkeley, CA: University of California Press.)

de Moivre, A. (1718) *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play*. London: W. Pearson. (2nd edition, 1738; 3rd edition, 1756. Reprinted 1967, New York, NY: Chelsea.)

De Morgan, Augustus (1845) Theory of probabilities. In *Encyclopedia Metropolitana*. Vol. 2: Pure Mathematics. pp. 393–490. London: B. Fellowes et al.

De Morgan, Augustus (1847) *Formal Logic: or, the Calculus of Inference, Necessary and Probable*. London: Taylor and Walton.

Diaconis, P. and Freedman, D. (1980) De Finetti's generalizations of exchangeability. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability*. Vol. 2. pp. 233–49. Berkeley and Los Angeles, CA: University of California Press.

Diaconis, P. and Freedman, D. (1985) Partial exchangeability and sufficiency. In Ghosh, J. K. and Roy, J. (eds.) *Statistics: Applications and New Directions*. Proceedings of the Indian Statistical Institute Golden Jubilee International Conference. pp. 205–36. Calcutta: Indian Statistical Institute,.

Diaconis, P. and Zabell, S. L. (1982) Updating subjective probability. *Journal of the American Statistical Association*. 77. pp. 822–30.

Dodds, E. R. (1951) *The Greeks and the Irrational*. Berkeley and Los Angeles, CA: University of California Press.,

Edwards, A. W. F. (1978) Commentary on the arguments of Thomas Bayes. *Scandinavian Journal of Statistics*. 5. pp. 116–18.
Ellis, R. L. (1843) On the foundations of the theory of probabilities. *Transactions of the Cambridge Philosophical Society*. 8. pp. 1–6.
Engel, Eduardo M. R. A. (1992) *A Road to Randomness in Physical Systems*. Lecture Notes in Statistics. Vol 71. New York, NY: Springer.
Falk, R. (1992) A closer look at the probabilities of the notorious three prisoners. *Cognition*. 43. pp. 197–223.
Feller, W. (1968) *An Introduction to Mathematical Probability*. 3rd edition, New York, NY: Wiley.
Fraser, D. A. S.(1968) *The Structure of Inference*. New York, NY: Wiley.
Gillies, D. (1987) Was Bayes a Bayesian? *Historia Mathematica*. 14. pp. 325–46.
Good, I. J. (1950) *Probability and the Weighing of Evidence*. New York, NY: Hafner Press.
Good, I. J. (1959) Kinds of probability. *Science*. 129. pp. 443–7.
Good, I. J. (1965) *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*. Cambridge, MA: MIT Press.
Good, I. J. (1971) 46656 varieties of Bayesians. *American Statistician*. 25. pp. 62–3.
Goodman, N. (1946) A query on confirmation. *Journal of Philosophy*. 43. pp. 383–5.
Goodman, N. (1979) *Fact, Fiction, and Forecast*. 3rd edition. Indianapolis, IN: Hackett,.
Gower, B. (1982) Astronomy and probability: Forbes versus Michell on the distribution of the stars. *Annals of Science*. 39. pp. 145–60.
Hacking, I. (1971a) Equipossibility theories of probability. *British Journal for the Philosophy of Science*. 22. pp. 339–55.
Hacking, I. (1971b) James Bernoulli's 'Art of Conjecturing'. *British Journal for the Philosophy of Science*. 22. pp. 209–29.
Hacking, I. (1975) *The Emergence of Probability*. Cambridge: Cambridge University Press.
Hald, Anders (1998) *A History of Mathematical Statistics from 1750 to 1930*. New York, NY: Wiley.
Heidelberger, M. (2001) Origins of the logical theory of probability: von Kries, Wittgenstein, Waismann. *International Studies in the Philosophy of Science*. 15. pp. 177–88.
Hill, B. (1968) Posterior distribution of percentiles: Bayes's theorem for sampling from a finite population. *Journal of the American Statistical Association*. 63. pp. 677–91.
Hintikka, J. (1966) A two-dimensional continuum of inductive methods. In Hintikka, J. and Suppes, P. (eds.) *Aspects of Inductive Logic*. pp. 113–32. Amsterdam: North-Holland.
Hintikka, J. and Niiniluoto, I. (1980) An axiomatic foundation for the logic of inductive generalization. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability*. Vol. 2. pp. 157–81. Berkeley, CA: University of California Press.
Hoppe, F. (1984) Polya-like urns and the Ewens sampling formula. *Journal of Mathematical Biology*. 20. pp. 91–4.
Howson, C. and Urbach, P. (2006) *Scientific Reasoning: The Bayesian Approach*. 3rd edition. Chicago and La Salle, IL: Open Court Press.
Jaynes, E. T. (2003) *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press.
Jeffrey, Richard C. (ed.) (1980) *Studies in Inductive Logic and Probability*. Vol. 2. Berkeley and Los Angeles, CA: University of California Press.
Jeffrey, Richard C. (1983) *The Logic of Decision*. 2nd edition. Chicago, IL: University of Chicago Press.

Jeffrey, Richard C. (1988) Conditioning, kinematics, and exchangeability. In Harper, W. L. and Skyrms, B. (eds.) *Causation, Chance, and Credence*. Vol. 1. pp. 221–55. Dordrecht: Kluwer.

Jeffreys, Harold (1939) *The Theory of Probability*. Oxford: Clarendon Press. (2nd edition 1958, 3rd edition 1961.)

Jevons, W. Stanley (1877) *Principles of Science: A Treatise on Logic and Scientific Method*. 2nd edition. London: Macmillan.

Johnson, W. E. (1924) *Logic, Part III: The Logical Foundations of Science*. Cambridge: Cambridge University Press.

Johnson, W. E. (1932) Probability: The deductive and inductive problems. *Mind*. 41. pp. 409–23.

Katz, Bernard D. and Olin, Doris (2007) A tale of two envelopes. *Mind*. 116. pp. 903–26.

Keynes, J. M. (1921/1973) *A Treatise on Probability*. In *The Collected Writings of John Maynard Keynes*. Vol. VIII. Cambridge: Cambridge University Press. (Originally published London, Macmillan.)

Kirk, G. S. and Raven, J. E. (1957) *The Presocratic Philosophers: A Critical History with a Selection of Texts*. Cambridge: Cambridge University Press.

Kuipers, Theo A. F. (1978) *Studies in Inductive Probability and Rational Expectation*. Dordrecht: Reidel,.

Lane, David A. and Sudderth, William D. (1978) Diffuse models for sampling and predictive inference. *The Annals of Statistics*. 6. pp. 318–36.

Laplace, Pierre Simon, Marquis de (1814/1995) *Essai philosophique sur les probabilités*. Paris: Courcier. (English translation, Dale, Andrew I. *Philosophical Essay on Probabilities*. Sources in the History of Mathematics and Physical Sciences. Vol. 13, New York, NY: Springer.)

Laplace, Pierre Simon, Marquis de (1812/1886) *Théorie analytique sur les probabilités*. In *Oeuvres complètes de Laplace*. Vol. 7. Paris: Gauthier-Villars. (Originally published Paris: Courcier.)

Lehmann, E. L. and Cassella, G. (1998) *Theory of Point Estimation*. 2nd edition. New York, NY: Springer.

Owen, G. E. L. (1966) Plato and Parmenides on the timeless present. *The Monist*. 50. pp. 317–40.

Poincaré, Henri (1896) *Calcul des Probabilités*. Paris: Gauthier-Villars.

Poincaré, Henri (1902/1905) *Science and Hypothesis*. Translated from the French by G. B. Halsted. New York, NY: The Science Press.

Poisson, Siméon-Denis (1837) *Recherches sur la probabilité des jugements en matières criminelles et matière civile*. Paris: Bachelier.

Ramsey, F. P. (1926/1964) Truth and probability. In Kyburg, H. E. and Smokler, H. E. (eds.) *Studies in Subjective Probability*. pp. 61–92. New York, NY: Wiley. (First read before the Cambridge Moral Sciences Club).

Rosenhouse, J. (2009) *The Monty Hall Problem*. Oxford: Oxford University Press.

Rosenthal, J. (2010) The natural-range conception of probability. In Ernst, G. and Hüttemann, A. (eds.) *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*. pp. 71–91. Cambridge: Cambridge University Press.

Shoesmith, E. (1985) Nicholas Bernoulli and the argument for divine providence. *International Statistical Review*. 53. pp. 255–9.

Shoesmith, E. (1987) The continental controversy over Arbuthnot's argument for divine providence. *Historia Mathematica*. 14. pp. 133–46.

Skyrms, B. (1991) Carnapian inductive logic for Markov chains. *Erkenntnis*. 35. pp. 439–60.

Skyrms, B. (1993) Analogy by similarity in hypercarnapian inductive logic. In Massey, G. J., Earman, J., Janis, A. I., and Rescher, N. (eds.) *Philosophical Problems of the Internal and External Worlds: Essays Concerning the Philosophy of Adolf Grünbaum*. pp. 273–82. Pittsburgh, PA: Pittsburgh University Press.

Stalker, D. (1994) *Grue! The New Riddle of Induction*. Chicago: Open Court.

Stigler, Stephen M. (1982) Thomas Bayes's Bayesian inference. *Journal of the Royal Statistical Society*. Series A. 145. pp. 250–8.

Stigler, Stephen M. (1986) *The History of Statistics*. Harvard University Press.

Strevens, M. (2011) Probability out of determinism. In Beisbart, J. C. and Hartmann, S. (eds.) *Probabilities In Physics*. pp. 339–64. Oxford: Oxford University Press.

Strong, John V. (1976) The infinite ballot box of nature: De Morgan, Boole, and Jevons on probability and the logic of induction. In *PSA 1976: Proceedings of the Philosophy of Science Association*. Vol. 1. pp. 197–211.

Swijtink, Zeno G. (1986) D'Alembert and the maturity of chances. *Studies in History and Philosophy of Science*. 17. pp. 327–49.

van Fraassen, Bas (1989) *Laws and Symmetry*. Oxford: Oxford University Press.

Venn, J. (1888) *The Logic of Chance*. 3rd edition. London: Macmillan.

von Kries, Johannes (1886) *Die Principien der Wahrscheinlichkeitsrechnung*. Tubingen: Mohr.

von Plato, Jan (1983) The method of arbitrary functions. *British Journal for the Philosophy of Science*. 34. pp. 37–47.

Whitworth, W. A. (1901) *Choice and Chance*. 5th edition. Cambridge: Deighton Bell.

Zabell, S. L. (1982) W. E. Johnson's "sufficientness postulate". *Annals of Statistics*. 10. pp. 1091–9.

Zabell, S. L. (1988) Symmetry and its discontents. In Harper, W. L. and Skyrms, B. (eds.) *Causation, Chance, and Credence*. Vol. 1. pp. 155–90. Dordrecht: Kluwer.

Zabell, S. L. (1989a) R. A. Fisher on the history of inverse probability. *Statistical Science*. 4. 3. pp. 247–56.

Zabell, S. L. (1989b) The rule of succession. *Erkenntnis*, 31. pp. 283–321.

Zabell, S. L. (1991) Ramsey, truth, and probability. *Theoria*. 57. pp. 211–38.

Zabell, S. L. (1992) Predicting the unpredictable. *Synthese*. 90. pp. 205–32.

Zabell, S. L. (1995) Characterizing Markov exchangeable sequences. *Journal of Theoretical Probability*. 8. pp.175–178.

Zabell, S. L. (1996) Confirming universal generalizations. *Erkenntnis*. 45. pp. 267–83.

Zabell, S. L. (1997) The continuum of inductive methods revisited. In Earman, J. and Norton, J. D. (eds.) *The Cosmos of Science: Essays of Exploration. Pittsburg-Konstanz Series in the Philosophy and History of Science*. pp. 351–85. Pittsburgh, PA: University of Pittsburgh Press/Universitätsverlag Konstanz.

Zabell, S. L. (2007) Carnap on probability and induction. In Friedman, M. and Creath, R. (eds.) *The Cambridge Companion to Carnap*. pp. 273–94. Cambridge: Cambridge University Press.

Zabell, S. L. (2011) Carnap on probability and induction. In Thagard, P. and Woods, J. (eds.) *Handbook of The History of Logic*. pp. 265–309. San Diego, CA: North Holland.

CHAPTER 16

FREQUENTISM

ADAM LA CAZE

16.1 Introduction

Axiomatic treatments of probability leave the interpretation of probability open. Kolmogorov (1950), in an early and much-discussed treatment, defines the relations between sets, fields, and probabilities, where "probability" can be given a classical, logical, frequency, propensity, best systems, or subjective interpretation.[1] This chapter will focus on the frequency interpretation. The frequency interpretation identifies probabilities with frequencies, either finite, observable frequencies, or limiting relative frequencies in infinitely repeated trials. Many areas of scientific practice rely extensively on probabilities interpreted as frequencies, specifically limiting relative frequencies. Key philosophically inclined accounts of frequentist probability include Venn (1888), von Mises (1957), and Reichenbach (1949/1971). More recent discussion tends to focus on frequentist statistical approaches, including for instance the account provided by Mayo (1996).

This chapter introduces the frequency interpretation of probability and the main criticisms leveled against it. While there is no denying that frequentist probabilities play a central role in much of science, there are straightforward (and some less-than-straightforward) problems with *identifying* probabilities with frequencies (Fine 1973; Hájek 1997, 2009). The problems with frequentism as an interpretation of probability raise interesting questions about the wide use of frequentist statistics in science. If frequentism is, at best, a flawed interpretation of probability, how can it play such an important role in science? A response to this question relies on an adequate distinction between the assessment of an interpretation as an analysis of probability and the use of methods that provide a frequentist model of probability.[2]

[1] See Lyon's "Kolmogorov's Axiomatization and its Discontents" (8) in this volume, as well as the individual chapters on specific interpretations in this volume. Additionally, see Hájek (2012).

[2] For a discussion of frequentist *methods* in comparison to Bayesian methods of statistical inference, see Sprenger's "Bayesianism vs. Frequentism in Statistical Inference" (18) in this volume.

16.2 What is the Frequency Interpretation of Probability?

16.2.1 Finite Frequentism

What is the probability of a woman suffering one or more heart attacks between the ages of 50 and 55? One direct method for arriving at this probability is to follow a sample of women aged 50 for 5 years and count the number who suffer one or more heart attacks. Imagine that the frequency of women suffering one or more heart attacks in the sample over the 5-year period is 3%. Providing the sample is large and is taken to adequately represent the population, it seems the observed frequency provides important information about the probability of interest.

Finite frequentism goes further and *defines* probability as the observed frequency. Finite frequentism, following the terminology of Hájek (1997: p. 212), defines the probability of attribute B in a finite reference class A as the relative frequency of actual occurrences of B within A. In notation form (following Reichenbach's (1949/1971) notation):

$$P(A,B) = F^n(A,B) = \frac{N^n(A \cap B)}{N^n(A)}$$

Here, $P(A,B)$ represents the probability of attribute B in reference class A, $F^n(A, B)$ represents the relative frequency of B in A, $N^n(A \cap B)$ represents the number of members of A with attribute B in the finite sample of size n, and $N^n(A)$ represents the number of members of the sample belonging to the reference class (in this case the reference class is defined by having attribute A, so $N^n(A) = n$). In the example, a first heart attack is attribute B, and the sample of women aged 50–55 is the reference class A.

It is hard to find supporters of finite frequentism from a philosophical perspective,[3] but it is often introduced in practical statistical texts as a viable interpretation of probability. Here is an example from a very good text on epidemiology (Rothman, Greenland, and Lash 2008: p. 10).

> The term *probability* has multiple meanings. One is that it is the relative frequency of an event…
>
> When one says "the probability of death in vehicular accidents when traveling >120 km/h is high," one means that the proportion of accidents that end with deaths is higher when they involve vehicles traveling >120 km/h than when they involve vehicles traveling at lower speeds (frequency usage).

Identifying probability with actual frequencies in finite samples has a number of well-recognized problems.

[3] One way to extend the view is to consider *possible* occurrences of attribute B within reference class A. When the reference class is finite, we might wish to call this finite hypothetical frequentism—more on hypothetical frequentism in the next section. Sklar (1973) provides a defense of finite hypothetical frequentism; for a critique, see Levi (1973).

The most famous problem for finite frequentism is *the problem of the single case*. According to finite frequentism all single-case events automatically have the probability 0 or 1. Consider a coin that is only tossed once and comes up Heads. It appears that the probability of Heads may be intermediate, but the finite frequentist is unable to say this. This goes against some strong intuitions about probability. A form of this problem remains in larger finite sequences. As the number of trials increases, the value of the denominator influences the frequency in a way that has nothing to do with probability. Consider the probability of Heads for a fair coin in a reference class of 1001 tosses. Despite the stipulation that the coin is fair, and the relatively large number of tosses, the probability of Heads can't be half according to the finite frequentist, and we know this *a priori*.

Identifying probability with finite frequencies posits too close a relation between actual outcomes in a finite sequence and probability. While frequencies are important to probability, describing the relationship between the two is no easy task.

16.2.2 Hypothetical Frequentism

Hypothetical frequentism avoids the problem of the single case by identifying the probability of an outcome with its relative frequency in a sufficiently large reference class. Specifically, the hypothetical frequentist suggests that the probability of a woman suffering one or more heart attacks between the ages of 50 and 55 is the *limiting relative frequency* of women suffering first heart attacks in an indefinitely large sample of the population. Likewise, for outcomes of an experimental process: for a simple experiment with a binary outcome, such as a coin toss, the hypothetical frequentist identifies the probability of Heads for the coin with the limiting relative frequency of Heads in an infinite sequence of trials.

The key philosophical proponents of hypothetical frequentism—Venn, von Mises and Reichenbach—articulate hypothetical frequentism in different ways, but there are similarities in their overarching view. Venn (1888: p. 11, italics in the original) identifies probability with limiting relative frequencies and makes an empirical claim about the existence of limiting relative frequencies:

> It will easily be seen that in every [series under discussion] there is a mixture of similarity and dissimilarity; there is a series of events which have a certain number of features or attributes in common,—without this they would not be classed together. But there is also a distinction existing amongst them; a certain number of other attributes are to be found in some and are not to be found in others. In other words, the individuals which form the series are compound, each being made up of a collection of things or attributes; some of these things exist in all the members of the series, others are found in some only. So far there is nothing peculiar to the science of Probability; that in which the distinctive characteristic consists is this;—that the occasional attributes, as distinguished from the permanent, are found on an extended examination to tend to exist *in a certain definite proportion of the whole number of cases*. We cannot tell in any given instance whether they will be found or not, but as we go on examining more cases we find a growing uniformity. We find that the proportion of instances in which they are found to instances in which they are wanting, is gradually subject to less and less comparative variation, and approaches continually towards some apparently fixed value.

The basic idea is that in a number of situations, and especially in standard games of chance, the relative frequency of an outcome tends to a specific proportion as the number of trials

increases. For example: in a large series of die rolls we expect there to be a tendency for the proportion of rolls with a "five" facing up to approach 1 in 6. The empirical claim Venn makes regarding the existence of limiting relative frequencies is called the *empirical law of large numbers* (and is discussed further in a moment). One of the challenges Venn faces is how to define the kinds of series in which the empirical law of large numbers is expected to hold. Venn (1888: p. 95) defines *ideal* sequences as those in which "the conditions of production, and consequently the laws of statistical occurrence, may be practically regarded as absolutely fixed" —such as in many standard games of chance. It is in ideal sequences that Venn suggests that the empirical law of large numbers holds. *Applying* the theory of probability to sequences in which the conditions of production cannot be regarded as fixed requires an assumption. Venn suggests:

> …to make the theory of the subject tenable, we have to really substitute one of this kind [ideal sequences] for one of the less perfect ones of the other class, when these latter are under treatment.
>
> (Venn 1888: p. 95).

The key contribution made by von Mises and Reichenbach is to formally specify the requirements on the sequences of experiments or observations that provide probabilities. The formal apparatus by which they achieve this differs. Von Mises' approach has been particularly influential, especially with respect to his approach to defining a random sequence (see Eagle's chapter "Probability and Randomness" (21) in this volume, and Eagle 2012). Von Mises restricts probability to particular types of infinite sequences, that is, those sequences that form *collectives*.

> A collective appropriate for the application of the theory of probability must fulfill two conditions. First, the relative frequencies of the attributes must possess limiting values. Second, these limiting values must remain the same in all partial sequences which may be selected from the original one in an arbitrary way. … The only essential condition is that the question whether or not a certain member of the original sequence belongs to the selected partial sequence should be settled *independently of the result* of the corresponding observation, i.e. before anything is known about this result. We shall call a selection of this kind a *place selection*.
>
> (von Mises 1957: pp. 24–5).

The informal idea behind a place selection is that it is impossible to construct a system whereby a gambler could use observed outcomes, or the location of values within a sequence, to bet successfully on unobserved outcomes. Place selections include: "select every odd member of original sequence" and "select the outcome following a run of four 'successes'". With collectives so defined, von Mises (1957: p. 15) defines probability:

> We will say that a collective is a mass phenomenon or a repetitive event, or, simply, a long sequence of observations for which there are sufficient reasons to believe that the relative frequency of the observed attribute would tend to a fixed limit if the observations were indefinitely continued. This limit will be called *the probability of the attribute considered within the given collective*.

There is no single-case (or indeed finite-case) probability for von Mises: no collective, no probability.

Reichenbach (1949/1971) provides an axiomatic account of probability (independent of Kolmogorov), and then argues for a hypothetical frequentist interpretation of this axiomatization that, in broad terms, shares many features of von Mises' account.

> If for a sequence pair $x_i y_i$ the relative frequency $F^n(A,B)$ goes toward a limit p for $n \to \infty$, the limit p is called the probability from A to B within the sequence pair.
>
> (Reichenbach 1949/1971: p. 69)

Here, x_i and y_i are sequences of elements (event-tokens). Some x_i belong to class A and some y_i belong to class B. Perhaps x_i is a sequence of tosses of several coins, with class A referring to the tosses of the specific coin we are interested in, and y_i including the outcomes of the various coin tosses. The sequence pair, $x_i y_i$, corresponds to the x_i that belong to A and y_i that simultaneously belong to B. Prior to determining the probability, the elements of the classes are put into one-to-one correspondence and ordered (Reichenbach 1949/1971: p. 48). Reichenbach (1949/1971: pp. 131–74) provides *axioms of order* to define the requirements of the sequence. Keeping these considerations in mind, and in the notation used earlier, Reichenbach's definition becomes:

$$P(A,B) = \lim_{n \to \infty} F^n(A,B)$$

Von Mises and Reichenbach differ in the requirements they place on sequences that provide probabilities (as well as how appropriate sequences are determined). Von Mises' condition of randomness places greater restrictions on the class of sequences. If any place selection can be made that changes the limiting relative frequency of an attribute in a subsequence, the original sequence is not a collective and the limiting relative frequency of the attribute is not a probability. Two criticisms of von Mises' account of a collective arise. The first is that it is not possible to provide a constructive proof for the existence of a collective.[4] The response to this problem favored by Reichenbach (and a number of others) is to restrict the class of place selections. Restricting the class of place selections permits constructive proofs for the existence of such sequences. Von Mises rejects this approach, and instead prefers to accept the "abstract 'logical' existence" of the collective. His worry is that the restriction of place selections arbitrarily excludes place selections that are not otherwise troubling (von Mises 1957: pp. 89–90). The second problem is that von Mises' characterization of place selections is mathematically inexact. It is possible to define a function that both meets von Mises' requirements for a place selection and identifies "successes" in an arbitrary sequence of "successes" and "failures" (Eagle 2012). Church (1940) employs the notion of "effective calculability" to render von Mises' characterization of place selections mathematically rigorous. In an added benefit of this formulation, Church (1940: pp. 134–5) goes on to show that existence proofs for random sequences so defined are necessarily non-constructive.

Von Mises and Reichenbach need to provide details of how *actual* sequences of observations or repeated trials relate to their formal characterization of probability. Venn's approach, as discussed above, is to "substitute" the sequence of trials under consideration with *idealized* series in which the "conditions of production" of the sequence are "regarded as absolutely fixed" (Venn 1885: p. 95). Von Mises (1957: p. 145) and Reichenbach (1949/1971:

[4] For discussion, see von Mises (1957: pp. 87–103) and Reichenbach (1949/1971: pp. 146–51).

p. 348) take a very similar approach—as do contemporary frequentist statisticians (see Cox 2006: p. 65). Von Mises, for example:

> An exact identity between theoretical premises and real conditions is not required, but only a similarity which makes a successful application of the theory to empirical data possible. The question that interests us is, therefore, what can be achieved by practical application of the theory of probability founded on the abstract concept of a collective?

Hypothetical frequentism gains traction on empirical probability problems to the extent that an empirical problem resembles, or can be made to resemble, the idealized scenarios in which the hypothetical frequentist definition of probability applies.

In application, the probability provided by the hypothetical frequentist is a counterfactual. Hypothetical frequentists identify the probability of an attribute with the limit of the relative frequency of an attribute that *would* be reached *were* we to repeat the trials indefinitely (and the sequence of trials is such that it meets, or can be assumed to meet, the formal requirements set out by the specific account). Hájek (2009: p. 212) provides a counterfactual definition of hypothetical frequentism:

> The probability of an attribute A in a reference class B is p if and only if the limit of the relative frequency of occurrences of A within B would be p if B were infinite.

This definition is consistent with the account of hypothetical frequentist probabilities provided by Venn, von Mises, and Reichenbach and provides a useful position against which criticisms can be considered.

16.2.3 Hypothetical Frequentism and the Law of Large Numbers

Given the importance of the *empirical law of large numbers* to hypothetical frequentism, it is important to clarify the nature of the link between this empirical claim and defining probability in terms of limiting relative frequencies. In particular, following von Mises (1957: pp. 104–15), it is important to distinguish the *empirical law of large numbers* from the mathematical theorems that go under the same name. The theorems, first discussed by Bernoulli and contemporaries, refer to the mathematical properties of particular sequences of numbers. The empirical law of large numbers refers to the claim that these properties are apparent in *actual* repetitions of certain types of trials. The mathematical theorems are briefly introduced, and then the empirical law of large numbers is discussed.

The *strong law of large numbers* can be formulated in terms of Bernoulli trials.[5] Consider an indefinite sequence of outcomes from a trial that has two possible outcomes, success or failure. An example is drawing and replacing marbles from an urn containing an unknown proportion of black and white marbles. Drawing a black marble is labeled "success". The outcome of each trial is independent and the probability of success is constant. x_i is the occurrence of success, n is the number of trials, p is the probability of success on any trial, and P is labeled the "meta-probability". The strong law of large numbers asserts that, with

[5] This follows the presentation of Fine (1973: p. 95)

meta-probability 1, the proportion of successes in n trials approaches the probability of success, p, as n approaches infinity.

$$P\left(\lim_{n\to\infty}\frac{1}{n}\sum_{i}^{n}x_i = p\right) = 1$$

Note the two probability statements. The first, p, refers to the probability of success, which is equivalent to the proportion of black marbles in the urn. The second, P, refers to the probability measure given to the sets of sequences of infinite draws in which the proportion of successes converges to p.

To say the proportion of successes will converge to *p with probability 1* is to say that the proportion of successes *almost surely* converges to *p*. The terms "with probability 1" and "almost sure convergence" are defined within measure theory. The meta-probability is given measure 1. Importantly, there exist sequences in which the proportion does not converge to *p*. *Success on every draw* is a possible sequence within the sample space. The existence of such sequences does not provide a contradiction to the strong law of large numbers. This sequence, along with infinitely many other "poorly behaved" sequences (i.e. sequences that do not converge to the probability in the limit), is given the probability measure 0.[6]

The strong law of large numbers, by itself, neither supports nor refutes the frequentist interpretation of probability. Bernoulli and Poisson formulated the theorems in terms of the classical interpretation of probability (see von Mises 1957: pp. 104–16). The frequentist needs to re-interpret probability in terms of frequencies and explicitly link the components of the mathematical theorems with actual experiments and experimental outcomes. The *empirical law of large numbers* performs this role. The empirical law of large numbers suggests that in sequences of outcomes of actual suitably selected experiments, the relative frequency of a specific attribute *almost surely* tends towards a limit. One way to define "suitably selected" is that the repetitive process under investigation is judged to provide outcomes that are independent and identically distributed across trials.[7] Importantly, "almost sure convergence" is also given a frequentist interpretation. Almost sure convergence is taken to provide a justification for assuming that the relative frequency of an attribute *would* converge to the probability in *actual* experiments *were* the experiment to be repeated indefinitely.

The expected outcomes of repetitions of standard games of chance are considered paradigmatic instantiations of the empirical law of large numbers. Some hypothetical frequentists refer to actual experimental setups of well-understood chancy processes repeated a very large number of times in which the observed outcomes approximate the law of large numbers to justify use of the empirical law of large numbers.[8] Nonetheless, believing the empirical law of large numbers will hold in any particular instance is always an article of faith. Two judgments need to be made. First, that the data-generating process under consideration provides a sequence of outcomes that meet the requirements of the specific frequentist account. Second, that the relative frequency of the attribute of interest in the

[6] See Fine (1973: pp. 95–7) for discussion.

[7] In many practical contexts this assumption is far from trivial.

[8] See Mayo (1996: pp. 165–8), for a discussion of Neyman's attention to the "empirical fact of long-run stability".

observed sequence of outcomes provides useful information on what the relative frequency would converge to in the limit. The *probability 1* guarantee that the relative frequency will converge on the probability is weaker than it first appears. Fine's (1973: p. 96) summary of the problem is apt:

> Probability 0 does not necessarily mean a small or negligible set of possibilities. … There are as "many" ways to get divergent relative-frequencies as there are to get convergent relative-frequencies and all the ways are equally probable.

The restrictions von Mises places on infinite sequences that provide probabilities means that his account suffers a slightly different version of this worry. For von Mises, probability can be discussed only in relation to a finite sequence on the assumption that the sequence is an initial segment of a collective. If the observed sequence is an initial segment of a collective, then provided the sequence is indefinitely extended, the relative frequency of any attribute will converge. The form of the problem that von Mises faces is that we are never in a position to know that any finite sequence is the initial segment of a collective.

16.3 Assessing Hypothetical Frequentism

An interpretation of probability can be assessed against a number of different criteria. Salmon (1967: pp. 63–5) suggests three criteria. First, an interpretation should be *admissible*: it should make true statements of the axioms of the formal system. Second, it must be possible (at least in principle) to *ascertain* the probabilities proposed by the interpretation. And third, knowledge of probabilities should have "practical predictive significance" (Salmon 1967: p. 64); that is, an interpretation should be judged according to its *applicability*.[9] An additional criterion, not mentioned by Salmon, is whether an interpretation *explicates commonsense notions of probability.*

Critics and proponents of hypothetical frequentism tend to differ in both the criteria they accept as appropriate for assessing the interpretation and the standard at which those criteria are assessed. Arguably, this is more a reflection of what is motivating the discussion than an explicit disagreement regarding appropriate criteria. The key criticisms of hypothetical frequentism considered in the following section are based on an assessment of hypothetical frequentism *as an interpretation*—that is, as to whether the interpretation provides an accurate analysis of probability. On this basis, each of the criteria is important, and the natural standard against which to assess these criteria is that of logical certainty.[10] By contrast, the primary motivation for Venn, von Mises, and Reichenbach in *defending* hypothetical frequentism was to better characterize the use of probability in science (as opposed to any other domain). Analyzing the way in which probability is used in science was

[9] Hájek (2012) offers additional criteria that are helpful in assessing interpretations of probability: *non-triviality, applicability to frequencies, applicability to rational belief, applicability to ampliative inference*, and *applicability to science.*

[10] For instance, when assessing *ascertainability*, the question is whether it is possible to ascertain hypothetical frequentist probabilities with certainty.

seen as a way of putting what we *mean* by probability statements on firm *objective* footing.[11] Von Mises (1957: p. 14) typifies this approach:

> There are in particular two points which I wish to emphasize: in the first place, the content of a concept is not derived from the meaning popularly given to a word, and it is therefore independent of current usage. Instead, the concept is first established and its boundaries are purposely circumscribed, and a word, as a suitable kind of label, is affixed later. In the second place, the value of a concept is not gauged by its correspondence with some usual group of notions, but only by its usefulness for further scientific development, and so, indirectly, for everyday affairs.

Given their motivation, the assessment of hypothetical frequentism by frequentists typically emphasizes its instrumental benefits. The focus is on what the *methods* achieve as opposed to what the interpretation explicates. Explication remains important, but the target has shifted. Instead of explicating commonsense probability concepts, the interpretation is seen to explicate the meaning of objective probability statements as they are used in the narrow domain of science. Intuitions about whether the interpretation explicates notions of probability outside the domain of science are considered irrelevant. This influences how the probability judgments that are required to arrive at hypothetical frequentist probabilities are viewed. Specifically, the justification of these assumptions is methodological; that is, rather than deductively proving propositions regarding the interpretation, an inferential rule is pragmatically justified on the basis of achieving a particular objective. On this view, hypothetical frequentism provides an idealized notion of probability that plays a role in operationalizing probability judgments in science.

Assessing hypothetical frequentism as an interpretation of probability shifts the focus. For instance, taking the view that an interpretation of probability should adequately capture commonsense probability concepts legitimizes the role of intuition in assessing the analysis of probability provided by hypothetical frequentism. To the extent that an interpretation is proffered as a single account, it needs to correctly explicate a minimal set of *essential* commonsense probability concepts. Some trade-off may need to be made between capturing commonly used probability concepts, and providing a philosophically precise analysis of probability. While opinions will differ on what constitutes the essential probability concepts, on this view, an interpretation should either explicate commonly used probability concepts or tell a compelling story about why the interpretation needs to eliminate some of these concepts from the formal account.

[11] Reichenbach (1949/1971: pp. 4–5), for instance:

> What can be the significance, for philosophical investigation, of a concept whose interpretation is vague and whose origin seems to be rooted in the inadequacy of human knowledge? … Only through the elaboration of the scientific theory of space and time, carried through in non-Euclidian geometry and the theory of relativity, did it become possible to uncover the ultimate nature of space-time concepts and to achieve a more profound understanding of their application to daily life. Thus we now have a better knowledge of what an architect means when he specifies lengths and widths in the plan of a building, and what a watchmaker does when he synchronizes a number of watches. … The probability concept, therefore, can be studied successfully only within the realm of its scientific application.

Venn (1888) and von Mises (1957) begin their discussion of probability with similar statements, and this sentiment is echoed in most frequentist texts (see, for instance, Feller 1968: pp. 1–6).

The different expectations on an interpretation of probability explain at least some of the diverging opinions on hypothetical frequentism. This chapter focuses on frequentism as an interpretation of probability, and so assesses hypothetical frequentism against each of the criteria listed above. Respecting the distinction between what is expected of an interpretation and methods developed on the basis of an interpretation is helpful in understanding why there can be compelling criticisms of hypothetical frequentism as an interpretation of probability while it retains an important role in science.

16.4 Problems with Hypothetical Frequentism

A number of criticisms have been made of hypothetical frequentism as an interpretation of probability (see especially: Fine 1973, Jeffrey 1992, and Hájek 2009). Hájek (2009) circles the hypothetical frequentist with fifteen counterarguments, and in doing so brings together many of the arguments that have been made. There are frequentist replies to some of the individual arguments that Hájek presents, but moves to avert one counterargument leave the frequentist exposed to one or more of the remaining 14. Sections 16.4.1 and 16.4.2 identify key problems for the identification of probability with limiting relative frequencies. Sections 16.4.3 and 16.4.4 address problems of ascertainability and applicability respectively. These problems can be considered both from the perspective of an interpretation and from the methods based on a specific interpretation of probability. Finally, Section 16.4.5 contains a consideration of the hypothetical frequentist's claims to objectivity.

16.4.1 Hypothetical Frequentism Provides a Poor *Analysis* of Probability

The hypothetical frequentist provides an answer to the question "What is probability?" with an analysis that has little relationship with what most people *mean* by the probability statements they make. Using a coin toss as an example, Hájek (2009: pp. 215–16) questions the sense of probability provided by the counterfactual that underpins hypothetical frequentism.

> We are supposed to imagine infinitely many results of tossing the coin: that is, a world in which coins are 'immortal', lasting forever, coin-tossers are immortal and never tire of tossing (or something similar, anyway), or else in which coin tosses can be completed in ever shorter intervals of time . . . In short, we are supposed to imagine *utterly bizarre* worlds . . . .
>
> (Hájek 2009: p. 215)

When stating that a specific coin has the probability of Heads of half, people are typically referring to their beliefs about the coin, their experience with this coin in a finite series of tosses, or their experience with similar-seeming coins. Assuming a suitable setup, and the acceptance of the frequentist's idealizations, a limiting relative frequency of the coin landing Heads of half may well be a *consequence* a particular chance set-up providing you are willing to accept the frequentist's idealizations, but this is different from asserting the hypothetical frequentist counterfactual as the correct (and only correct) *analysis*.

Of course, frequentists don't expect coins or coin-tossers to actually provide an infinite sequence of tosses in which the relative frequency can be determined. Rather, they suggest that infinite sequences should be viewed as an *idealization*—unobservable, but nonetheless an aid to understanding and assessing probability. But this abstraction doesn't resolve the problem of the hypothetical frequentist's analysis of probability. Hájek (2009: p. 217) responds:

> [Hypothetical frequentism] asserts a biconditional (one which presumably is supposed to be necessarily true). According to it, the coin in my pocket lands Heads with probability 1/2 *if and only if* a certain bizarre counterfactual about a very different coin is true. I am puzzled by the claim that the two sides of this biconditional (necessarily) have the same truth value; indeed, I flatly deny it. Is my puzzlement supposed to vanish when I am told that [hypothetical frequentism] involves an *idealization*?

The distinction between interpretation and method is central to this criticism. Hypothetical frequentist methods operationalize probability in a way that makes a range of scientific problems tractable. But a successful method need not be a successful analysis.[12] Judged as *analysis*, and more particularly a comprehensive analysis of probability, hypothetical frequentism is simply false.

16.4.2 The Relation Between Frequencies and Probabilities is not Identity

Hypothetical frequentism posits an identity relation between hypothetical frequencies and probability. But, as with actual relative frequencies, the relationship between hypothetical relative frequencies and probability can't be as close as identity: there are hypothetical relative frequencies that don't meet the requirements of the probability calculus, and there are probabilities that can't be understood in terms of hypothetical relative frequencies.

Kolmogorov's axiomatization of probability includes an additivity axiom.[13] Finite additivity holds that the probability of the union of a finite number of disjoint events is equal to the sum of the probabilities of all the events:

$$P(E_1 \cup E_2) = P(E_1) + P(E_2), \text{if } E_1 \cap E_2 = \emptyset.$$

Finite additivity has the uncontroversial consequence that the probability of rolling a "2", "4" or "6" in a single roll of a standard die is the sum of the probabilities of all these events, that is $P(\text{“2”} \cup \text{“4”} \cup \text{“6”}) = P(\text{“2”}) + P(\text{“4”}) + P(\text{“6”})$. In infinite sequences, the axiom of *countable additivity* holds that if $E_1, E_2, \ldots, E_i$ are mutually disjoint events within a Borel (σ-) field, F, on outcome space, K, then $P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$. Whether countable

[12] Hájek (1997: p. 226) provides a helpful analogy. Consider the biconditional: *X is a bachelor if and only if X ticks the "bachelor" box on a questionnaire*. Whatever the success of using this biconditional as a method for identifying bachelors, it clearly provides a false analysis of the concept "bachelor".

[13] The criticism as presented here assumes Kolmogorov's axiomatization of probability. While dominant, Kolmogorov's approach is not the only option for axiomatizing probability. See Lyon's "Kolmogorov's Axiomatization and its Discontents" (8) in this volume.

additivity should be an axiom of probability is controversial, but hypothetical frequentists are going to need countable additivity or something quite like it.[14] Van Fraassen (1980: p. 185) and Hájek (2009: pp. 229–30) use similar cases to illustrate hypothetical frequentism's violation of countable additivity.

The following scenario is Hájek's. Consider an infinite lottery, with tickets numbered with the positive integers. A_i = "ticket i is drawn". Consider also an infinite sequence of draws (with replacement). A possible sequence of outcomes that could arise from this setup is each ticket being drawn exactly once. The relative frequency of obtaining each ticket in the limit is zero, so, according to the hypothetical frequentist: $P(A_i) = 0$, for all i. Thus $\sum_{i=1}^{\infty} P(A_i) = 0$. But since there is a ticket selected on every draw, and each ticket is drawn, $A_1 \cup A_2 \cup A_3 \cup \ldots$ happens every time, and $P(\bigcup_{i=1}^{\infty} A_i) = 1$. The infinite lottery case demonstrates that hypothetical frequentism is not admissible.

Identifying probability with limiting relative frequencies afflicts probability with the complexity of infinite sequences. Not every phenomenon provides outcomes that will approach a limit in an infinite series of trials, and the limit is sensitive to aspects of the trials that should not influence probability. Hájek (2009) discusses infinite sequences with: (i) more than one limiting relative frequency, due to the lack of a unique ordering of the sequence; (ii) no limit, because the sequence oscillates between outcomes; (iii) limits that are possible, but that conflict with well-grounded probability judgments (think of a fair coin that happens to come up Heads on every toss); and (iv) a possible limit despite the outcome of interest falling in a "probability gap" (consider an infinitely thin dart landing in a non-measurable region of a dartboard).[15]

Hypothetical frequentists respond to worries such as these by making a number of methodological assumptions. For instance, the existence of a (single) limit is an explicit assumption in both Reichenbach's and von Mises' account. Reichenbach (1949/1971: pp. 52–3) calls this the *rule of existence.* Further assumptions are required to ensure that the hypothetical frequentist probability can be ascertained and/or applied to scientific questions. For instance, both Reichenbach and von Mises assume that the actual sequence of the experimental outcomes will be (or can be made to be) a *suitable sequence* for ascertaining probabilities. For many frequentist *methods*, the assumption of a suitable sequence is that the outcomes of trials are independently and identically distributed.

These assumptions are required to get the methods off the ground; the extent to which they can be defended from a methodological perspective will depend on the details of the setup and knowledge about the process under investigation. The assumptions are typically impossible to prove, or known to be false. The best that is hoped for is that the real distribution varies from the idealized distribution in a way that does not undermine the overall analysis. The high level of acceptance of frequentist methods in many areas of science is testament to the willingness of scientists to make and accept such judgments.

Finally, there are also cases in which there are probabilities that are not relative frequencies. Hypothetical frequentism restricts itself to instances in which it is possible to make sense of an observable repetitive structure, and, in particular, an observable

[14] Van Fraasen (1980: pp. 182–7) provides discussion on why infinite sequences, and thus countable additivity, are required.

[15] See Hájek (2009) sections 4, 5, 6, and 8 respectively.

repetitive structure in which the individual events can be considered independent and of a constant (though unknown) probability. There are many situations in which there is wide agreement that the hypothetical frequentist conception fails to apply. Some statements, as Hájek (2009: pp. 221–2) notes, are single-case events by necessity. An observable repetitive structure is not going be possible for universal generalizations or existential statements. And yet, scientists and others often wish to discuss the probability of the truth of a large-scale theory, say, the probability of evolutionary theory being true, or the probability of higher intelligence evolving in silicon-based life forms. Hypothetical frequentism is unable to make such one-off questions directly tractable. Once again, this is a problem for hypothetical frequentism as an interpretation of probability. As a basis for frequentist methods, proponents of hypothetical frequentism are able to accept that hypothetical frequentist probabilities are limited to situations in which a certain kind of repetitive structure is possible.

16.4.3 Hypothetical Frequencies are Strictly Unascertainable

To ascertain a hypothetical frequency with certainty we would need to observe an infinite number of trials. Assuming that a specific sequence of observations will converge to a limiting relative frequency, there is no guarantee that it will do so within the number of trials that will be observed. And if a relative frequency appears to have converged in a finite number of trials, it is always possible that the relative frequency diverges from this value in subsequent trials. These points are direct consequences of the mathematics of infinite sequences. The task for the frequentist is to justify inferring a (frequentist) probability from a relative frequency observed in a finite number of trials, and there is no deductively valid way to do this.

The frequentist's problem of ascertaining probabilities is a restatement of the problem of induction in frequentist form. Reichenbach's response to this problem is to provide a pragmatic justification of a methodological rule for inferring the value of a limiting relative frequency based on the relative frequency observed in a finite number of trials.[16] The inferential procedure is called *induction by enumeration* and his methodological rule, the *rule of induction*. The idea is that assuming a limiting relative frequency exists, the value of that limit can be inferred as finite data accumulates. As increasing amounts of data are collected the inferred value of the limit is updated. Reichenbach justifies his rule on the basis that (i) *assuming a limiting relative frequency exists,* repeated application of the methodological rule will converge on the limiting relative frequency; and (ii) there are no alternative inferential rules that can be proven to perform better. Reichenbach acknowledges that since it is not possible to apply the procedure long enough to be sure the limiting relative frequency is reached, the value of the limit is only ever posited or inferred—it is never proven.

Salmon (1967: pp. 84–108, 1991) identifies the key problem facing Reichenbach's approach: there are infinitely many rules that have the same convergence properties as induction by enumeration, but that provide diverging advice for inferring the value of the limit based on an observed finite sequence. These are *asymptotic rules*:

[16] Reichenbach (1949/1971) see especially chapters 9 and 11. Salmon (1967: pp. 84–108 and 1991) provide a critical discussion of Reichenbach's argument.

> If an observed initial section consisting of n members of a sequence of As contains m elements with the attribute B **POSIT THAT** the limit of the relative frequency of B in A lies within the interval $\left(\frac{m}{n} + c_n\right) \pm \delta$, where $c_n \to 0$ as $n \to \infty$.
>
> (Salmon 1991: p. 104)

Salmon (1991) surveys attempts to supplement Reichenbach's rule of induction in order to justify his particular asymptotic rule, but none raise the selection of the rule of induction above methodological stipulation.

16.4.4 Applying Hypothetical Frequencies and the Reference Class Problem

A key criterion for judging the success of any interpretation of probability worth the name is whether the interpretation provides probabilities that can be successfully applied to scientific problems. An essential feature in successful application of an interpretation of probability will be the predictive success of the probabilities provided by the account. In frequentist terms, the question becomes whether hypothetical relative frequencies predict outcomes in finite sequences. There are a number of problems of application for hypothetical frequentist probabilities. The first of these is how outcomes in finite sequences should be conceived by the hypothetical frequentist.

Consider the probability that Jill, aged 50, will suffer her first heart attack in the next five years. Von Mises (1957: p. 11) and Reichenbach (1949/1971: pp. 376–7) suggests that such single-case probability statements are meaningless.

> We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase 'probability of death', when it refers to a single person, has no meaning at all for us.
>
> (von Mises 1957: p. 11)

As Hájek (2007: p. 568) notes, given that hypothetical frequentists define probability in terms of an infinite reference class, finite-case probabilities present a similar problem for the hypothetical frequentist. The probability of a first heart attack in the next five years for the next thirty 50-year-old women a physician sees is as meaningless to the hypothetical frequentist as Jill's probability of a heart attack. So too is this probability for a reference class of 5 million women, or of 5 trillion, or of any finite number of women.

This is a problem for the strict applicability of hypothetical frequencies. There is no necessary link between the limiting relative frequency of an attribute in an infinite sequence and the relative frequency of that attribute observed in a finite sequence. The same problem arises regardless of the direction in which you wish to move—from a hypothetical frequentist probability to a relative frequency in a finite sequence, or vice versa. The problem of application is the challenge of ascertaining hypothetical frequentist probabilities in reverse. Fine (1973: p. 102) summarizes the problem:

> Even with the assurance that $\lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} x_i = p$, we are unable to say anything about $x_i, \ldots, x_n$. The events $\left\{ \lim_{n\to\infty} \frac{1}{n} \sum_{i}^{n} x_i = p \right\}$ and $\{x_i, \ldots, x_n\}$are independent—knowing the first occurs tells us nothing about the second.

Given the similarities between the problems of ascertaining and those of applying hypothetical frequentist probabilities, the best the frequentist can do is provide a similar methodological response to the problem of application. Essentially, the frequentist needs to assume that the relative frequency of the attribute in a finite sequence will approximate the probability. The most assurance the hypothetical frequentist can provide is that the resemblance should improve as the finite sequence gets larger.

Von Mises' and Reichenbach's approach to single- and finite-case probabilities is to locate the attribute of interest within an infinite reference class. Reichenbach (1949/1971: p. 375), for instance:

> . . . [T]here exists only one legitimate concept of probability, which refers to classes, and the pseudoconcept of probability of a single case must be replaced by a substitute constructed in terms of class probabilities.

Putting aside the applicability problem already discussed, the selection of an appropriate reference class poses additional challenges.

Consider again the probability that Jill, aged 50, will suffer her first attack in the next five years. Jill is a healthy Italian who, among other pursuits, enjoys running and deep-sea diving. Imagine that five years' worth of health data is available on an indefinitely large number of people. x_i contains the demographic and personal details about the people in the dataset; y_i contains information on each person's five-year health outcomes. To get a relative frequency it is necessary to select the members of x_i that belong to the reference class, call this class A, and the members of y_i that belong to the class of interest, class B. Once classes A and B are defined, so too is the hypothetical frequentist probability.

Let's assume that defining class B is unproblematic; there is a clear medical definition of "heart attack" and it is reported accurately in the dataset. How should the reference class A be defined? Jill has many attributes, and so is a member of many classes. She is a member of the class of 50-year-olds, she is also a member of the class of over-45s, as well as the class of runners, the class of deep-sea divers and the class of Italians. Should the reference class be broad, and include everything we know about Jill, or narrow, and include only those memberships that are thought to influence heart disease? Each reference class may provide a different probability. There appears to be no clear-cut answer to the question of which reference class should be used to judge Jill's chances of a heart attack. This is an example of the *reference class problem*. The reference class problem arises when the probability of an event may vary when the event is classified in two or more ways—each of which appears to be an equally legitimate way to classify the event. Hájek (2007: p. 565) provides a general formulation of the problem:

> Let X be a proposition. It seems that there is *one unconditional* probability of X; but all we find are *many conditional* probabilities of the form $P(X,$ given $A)$, $P(X,$ given $C)$, etc. that differ from each other. Moreover, we cannot recover $P(X)$ from these conditional probabilities by the law of total probability, since we likewise lack unconditional probabilities for A, B, C, etc. (and in any case A, B, C, etc. need not form a partition). Relativized to the condition A, X has one probability; relativized to the condition B, it has another; and so on. Yet none of the conditions stands out as being the *right* one.

While attempts have been made to resolve the reference class problem, none are ultimately successful (if we take success to be the establishment of a general method for arriving at a

single unarguable objective probability).[17] Importantly, Hájek (2007) shows that a version of the reference class problem arises for any interpretation of probability that provides a guide to action. A strict subjectivist can avoid the reference class problem, but only for as long as they refrain from incorporating empirical evidence into the evaluation of their probability.[18]

Hájek (2007: pp. 567–9) distinguishes two variants of the reference class problem—one metaphysical and one epistemological. The metaphysical variant raises questions about the nature of probability. It is typically assumed that it is meaningful to talk of the probability of an event, such as the probability of Jill's having a heart attack, as an unconditional probability. Kolmogorov's axiomatization treats unconditional probability as primitive; conditional probability is then defined in terms of unconditional probability. The pervasiveness of the reference class problem undermines this analysis. For this reason, and a number of others,[19] Hájek argues that conditional probabilities are primitive.

> We should accept that there are only *relative* probabilities out there—probabilities *conditional* on this condition or that ($P(X$, given $A)$, $P(X$, given $B)$, etc).
>
> (Hájek 2007: p. 583)

This resolves the metaphysical reference class problem by seeing it as an inescapable part of probability.

This leaves us with the epistemological reference class problem (Hájek 2007: pp. 583–4). Accepting conditional probability as primitive doesn't ease the difficulty of choosing which conditional probability should be used in a specific case. In the absence of a solution to the epistemological reference class problem, frequentists provide methodological advice. Salmon (1971: pp. 40–7), for instance, suggests that one should apply the *reference class rule*:[20] "choose the broadest homogeneous reference class to which the single event belongs". A "homogeneous reference class" is defined using von Mises' notion of a place selection. A homogeneous reference class for an attribute is a class in which there is no place selection that produces a subclass for which the probability of the attribute in the subclass differs from the probability of the attribute in the entire reference class. All possible place selections in a homogeneous reference class are said to be *statistically irrelevant* to the attribute in the reference class. Salmon's reference class rule seems to provide good methodological advice, but it fails to resolve the epistemological reference class problem because there is no assurance there will be a unique homogeneous reference class.

[17] See Hájek (2007: pp. 567–9) for discussion.

[18] As soon as the subjectivist identifies a reference class in which the probability of suffering a heart attack is viewed as equiprobable—following the approach outlined by de Finetti (1937/1980: p. 75)—they will be faced with the same problem as the frequentist: multiple reference classes in which the members are judged to be equally likely to suffer a heart attack, and thus, multiple probabilities.

[19] See Hájek (2003; 2007: pp. 581–3)

[20] Salmon's methodological advice is an improvement on Reichenbach's (1949/1971: p. 374) less-than-clear suggestion to consider "*the narrowest class for which reliable statistics can be compiled*". The problem, as Salmon (1971: p. 41) notes and Reichenbach (1949/1971: p. 375) acknowledges, is that it is not always clear how to apply this advice. Narrowing the reference class will typically reduce the reliability of the available statistics. It is not at all clear at what point selecting the "narrowest" reference class will produce statistics that are sufficiently "reliable".

16.4.5 Hypothetical Frequentism and Objectivity

Objectivity is taken to be a key feature of the frequentist interpretation of probability. However, the notion that frequentism is objective is typically taken as a starting premise. There is relatively little by way of direct analysis regarding what it means for frequentism to be an objective interpretation of probability. Subjective interpretations of probability identify probability—in one way or another—with the beliefs of individual agents. Frequentists identify probability with frequencies, and *not* an individual's beliefs. Does the frequentist's claim to objectivity amount to more than this distinction?

Von Mises takes a strong position regarding the objectivity of hypothetical frequentism. He views the limiting relative frequency of an attribute in a given collective as a *physical* property of the phenomenon of interest in a given setup.

> The probability of a 6 is a physical property of a given die and is a property analogous to its mass, specific heat, or electrical resistance. Similarly, for a given *pair of dice* (including of course the total setup) the probability of a 'double 6' is a characteristic property, a physical constant belonging to the experiment as a whole and comparable with all its other physical properties. The theory of probability is only concerned with relations existing between physical quantities of this kind.[21]
>
> (von Mises 1957: p. 14)

On von Mises' view, probability interpreted as hypothetical frequencies is *in the world* in the same way as many other properties measured in science are in the world. Just as an object has a mass whether or not it is measured, a coin has a probability whether or not it is tossed indefinitely.

Jeffrey (1992: p. 194) rejects von Mises' analogy.

> If one could and did toss the coin forever (without changing its physical characteristics) one would have brought such a sequence into physical existence, just as one would have brought an extra planet into existence by suitable godlike feats, if one were capable of them and carried them out. But in the real world, neither the sequence nor the planet exists, and the one is as far from having a limiting relative frequency of heads as the other is from having a mass.[22]

Jeffrey admits a point of disanalogy in this comparison. Whereas we would be silent on the planet's mass, we are usually happy to judge the coin's probability of landing Heads. However, this only provides further support for Jeffrey's position. Given that the infinite sequence of trials does not exist, any judgment about the coin's probability is based on factors *other* than the hypothetical frequency of Heads in an infinite number of tosses—the physical appearance of the coin, outcomes of previous tosses, and the like.

Hypothetical frequencies need not be physical to be objective. By definition, if a sequence of trials is part of a collective, then the relative frequency of an attribute will have a limiting

[21] It is worth noting that the physical property that von Mises is referring to is the limiting relative frequency of an attribute in a collective, and not *other* physical properties of the object (e.g. the physical dimensions or weight distribution of a die or a coin).

[22] Hájek's arguments against the hypothetical frequentist's biconditional also pick up on this point (Hájek 2009: pp. 217–20). While there is a fact of the matter about a fair coin's probability of landing Heads, namely half, there is *no* fact of the matter of what the relative frequency of Heads would be if the coin were tossed infinitely many times. Depending on the ordering of the sequence of tosses, there may be more than one limiting relative frequency or there may be no limiting relative frequency.

value, and this limiting relative frequency exists independently of any person's beliefs. Reichenbach and other frequentists would tell a different story, but the conclusion is similar. If a limiting relative frequency exists, it exists independently of any thinking subject. This is the sense in which hypothetical frequentist probabilities are objective, and it is broadly consistent with the notion of objectivity discussed in statistical texts. Barnett (1999: p. 70), for instance, suggests that hypothetical frequencies are objective because they are "divorced from any consideration of personal factors". But notice that this is true only *after* a model for the data-generating process has been specified (or assumed). As discussed in Section 2.2, to apply hypothetical frequencies in real-world scenarios assumptions are required regarding the sequence of outcomes the data-generating process would supply were it to be indefinitely repeated. For von Mises, you need to assume the sequence will form a collective; for Venn, an ideal sequence; for Reichenbach, a sequence that obeys the axioms of order. Scientists employing frequentist probabilities need to make a judgment that the data-generating processes providing the measured outcomes of the study are adequately modeled by one or more of these approaches to specifying the requirements on the expected sequence of outcomes. This is an important judgment. In some situations there will be broad agreement regarding the expected sequence of outcomes, and in some situations individual judgments will differ.

Hypothetical frequencies are not divorced from consideration of personal factors (including beliefs). All viable approaches to probability and statistics rely on judging, assuming and/or specifying a model for the data-generating process under investigation. The difference between frequentist and subjective approaches in terms of objectivity is more methodological. Once the model is adequately specified, the frequentist models residual uncertainty in terms of frequencies, whereas most subjectivist approaches model uncertainty in terms of beliefs.

16.5 Conclusion

There are compelling reasons not to identify probability with hypothetical frequencies in infinite sequences. Probabilities and hypothetical frequencies are linked, but they are not the same. This does not necessarily undermine the utility of methods developed on the basis of a hypothetical frequentist interpretation of probability. Arguments for or against specific methods relying on different interpretations of probability will be influenced by a range of factors, including the context and the advantages and disadvantages of the method in developing answers to the question at hand.

Acknowledgments

I am grateful to Alan Hájek for many helpful comments on this chapter.

References

Barnett, Vic (1999) *Comparative Statistical Inference*. Chichester: Wiley.

Church, Alonzo (1940) On the Concept of a Random Sequence. *Bulletin of the American Mathematical Society*. 46. 2. pp. 130–5.

Cox, D. R. (2006) *Principles of Statistical Inference*. Cambridge: Cambridge University Press.

de Finetti, Bruno (1937/1980) Foresight: Its Logical Laws, its Subjective Sources. In Kyburg, H. E. Jr. and Smokler, H. E. (eds.) *Studies in Subjective Probability.* 2nd edition. Huntington, NY: Robert E. Krieger Publishing Company. (Originally published in French.)

Eagle, Antony (2012) Chance versus Randomness. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring. [Online] Available from: http://plato.stanford.edu/archives/spr2012/entries/chance-randomness/. [Accessed 8 Oct 2015.]

Feller, William (1968) *An Introduction to Probability Theory and its Applications*. New York, NY: Wiley.

Fine, T. L. (1973) *Theories of Probability*. New York, NY: Academic Press.

Hájek, Alan (1997) "Mises Redux"—Redux: Fifteen Arguments Against Finite Frequentism. *Erkenntnis*. 45. pp. 209–27.

Hájek, Alan (2003) What Conditional Probability Could Not Be. *Synthese*. 137. 3. pp. 273–323.

Hájek, Alan (2007) The Reference Class Problem is Your Problem Too. *Synthese*. 156. pp. 563–85.

Hájek, Alan (2009) Fifteen Arguments Against Hypothetical Frequentism. *Erkenntnis*. 70. pp. 211–35.

Hájek, Alan (2012) Interpretations of Probability. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. Spring. [Online] Available from: http://plato.stanford.edu/archives/sum2012/entries/probability-interpret/. [Accessed 8 Oct 2015.]

Jeffrey, Richard (1992) Mises Redux. In *Probability and the Art of Judgment*. pp. 192–202. Cambridge: Cambridge University Press.

Kolmogorov, A. N. (1950) *Foundations of the Theory of Probability*. Oxford: Chelsea Publishing Co.

Levi, I. (1973) But Fair to Chance. *The Journal of Philosophy*. 70. 2. pp. 52–5.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.

Reichenbach, H. (1949/1971) *The Theory of Probability*. Repr. Berkeley, CA: University of California Press.

Rothman, Kenneth J., Greenland, Sander, and Lash, T. L. (2008) *Modern Epidemiology*. 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Salmon, Wesley C. (1967) *The Foundations of Scientific Inference*. Pittsburgh, PA: University of Pittsburgh Press.

Salmon, Wesley C. (1971) *Statistical Explanation and Statistical Relevance*. Pittsburgh, PA: University of Pittsburgh Press.

Salmon, Wesley C. (1991) Hans Reichenbach's Vindication of Induction. *Erkenntnis*. 35. pp. 99–122.

Sklar, L. (1973) Unfair to Frequencies. *The Journal of Philosophy*. 70. 2. pp. 41–52.

van Fraassen, Bas C. (1980) *The Scientific Image*. New York, NY: Oxford University Press.

Venn, John (1888) *The Logic of Chance*. London: Macmillan and Co.

von Mises, Richard (1957) *Probability, Statistics and Truth*. New York, NY: Macmillan.

CHAPTER 17

SUBJECTIVISM

LYLE ZYNDA

Subjectivism in probability theory is the view that probabilities are degrees of belief. Thus, when I make an assertion about probabilities, I may be reporting (or expressing) how confident I am that something is true; alternatively, I can refer to "your" probability that something is true, i.e., your degree of belief in it. The key thing is that a subjective probability is a feature of the psychological state of some specific person or other.

A strict subjectivist (such as de Finetti) would hold that this is the only or primary interpretation of probability; however, many believers in subjective probability (sometimes also called "personal probability") would allow that probability sometimes has other meanings, too, such as various sorts of objective chance (e.g., physical propensities), limiting long-run relative frequencies, etc. It is just that there are certain contexts—and important ones—in which probability doesn't make sense unless interpreted as degree of belief. For example, one-time events can be assigned a degree of belief: so, when bookmakers placed odds on whether Obama or Romney would win the 2012 US election, they were expressing their degrees of belief. One-time events such as particular elections are by definition not repeatable like coin tosses (one cannot run the 2012 election over and over and see how often Obama wins against Romney). Also, scientific theories are not repeatable events (they are not events at all), nor are they features of the physical world, like radium atoms' propensity to decay in a given period. Nonetheless, we judge theories as more or less likely, given the evidence at our disposal. Those who favor logical probability (such as Keynes and Carnap) understand probability in such contexts to be a relation of partial implication between statements, and would analyze evidential support as a measure of the extent to which our evidence implies a given theory. Ideally, this would be captured by some unique logical probability measure, or a limited class of such measures. Subjectivists, by contrast, take such relations to indicate the extent to which our evidence raises (or would raise, under certain conditions) our confidence in a theory, and hold that this can legitimately vary from person to person (within certain bounds). Indeed, legitimate disagreement (again, within certain bounds) is an essential feature of subjectivism in probability theory. My probabilities may not be yours.

Because subjectivists understand probability to be degree of belief, probability theory for them has both descriptive and normative aspects. First, there has to be a way to determine or "measure" one's actual degrees of belief (this depends on what degrees of belief *are*, an

issue we will discuss in subsequent sections); then, some degrees of belief are judged to be "rational" in some sense (and others not). Regarding the second issue, subjectivists all agree that degrees of belief ought to obey the laws of the probability calculus.[1] In particular, degrees of belief ought to be non-negative, with logical truths receiving maximal probability (the maximum by convention normally set to 1), and they must be *finitely additive*—i.e., for any two logically exclusive statements p and q, $pr(p \vee q) = pr(p) + pr(q)$. In addition, conditional probabilities $pr(p|q)$ (read "the probability of p given q") are constrained so that $pr(p|q) = pr(p \,\&\, q)/pr(q)$ whenever $pr(q) > 0$.[2] Beyond this, however, there are several points of *disagreement* among subjectivists over what formal conditions constrain legitimate ("rational") degree of belief. Many subjectivists would require that, in addition to being finitely additive, degrees of belief be *countably* additive as well—i.e., that whenever $\{p_i\}$ is a countable set of pairwise logically exclusive statements, $pr(p_1 \vee p_2 \vee \ldots) = \Sigma pr(p_i)$. Other rationality constraints are sometimes (though less commonly) added, such as reflection or regularity (defined below, and discussed in detail elsewhere in this volume, e.g., Kotzen (29)), or the Principal Principle (see Schwarz, also in this volume (20)). However, none of these additional constraints has the universal assent of all subjectivists. Thus, we may take as the minimal normative core of subjectivism the thesis that degrees of belief ought to range from 0 to 1, with logical truths receiving maximal probability, and that they ought to be finitely additive as well.

Besides these "synchronic" constraints, which describe how degrees of belief are related to each other at a single time, subjectivists also agree to certain "diachronic" constraints about how one's degrees of belief should change over time, in response to evidence.[3] Suppose that one learns that e, and nothing more, between times t and $t^\star$. Then one's probability for any statement p at $t^\star$, $pr_{t*}(p)$ (called one's "posterior" probability), should be such that $pr_{t*}(p) = pr_t(p|e)$, where pr_t is one's "prior" probability function at t. The notion of priority is a relative one, since as learning proceeds, one's old "posterior" probabilities become one's new "prior" probabilities, relative to evidence not yet obtained. Thus, upon learning evidence $e^\star$ after e, and nothing more, one would update pr_{t*} to $pr_{t**} = pr_{t*}(\text{-}|e^*)$, and so on. The basic process of updating just described is called *conditionalization* (or alternatively, *conditioning*). Other forms of updating that apply in contexts other than those in which what is learned can be expressed in a single proposition e include *Jeffrey conditioning*. Jeffrey (1965) generalized the idea of conditionalization, so that if pr and $pr^\star$ are one's probabilities now and at some later time, and $\{e_i\}$ is a partition of the probability space such that $pr(p|e_i) = pr^*(p|e_i)$ for all p and e_i, then $pr^*(p) = \Sigma pr(p|e_i)pr^*(e_i)$. Thus, one can think of Jeffrey conditioning (or "probability kinematics," as he preferred to call it) as the result of redistributing probabilities across a partition while leaving the relative weights inside each e_i unchanged.[4] Jeffrey

[1] That is, when they are precise. Subjective probabilities may be vague, and in such cases ought to be subject to somewhat more complicated constraints. (See Fabio Cozman's chapter "Imprecise Probabilities," (14) in this volume, for further discussion.)

[2] Typically, this is stated as a "definition" of conditional probability, but this cannot be strictly correct, as Hájek (2003) shows; e.g., conditional probabilities $pr(p|q)$ can have well-defined values in certain cases where $pr(q) = 0$.

[3] Normally it is assumed that *only* changes of this sort, i.e., those made in response to evidence via general updating rules, are "rational."

[4] It is worth pointing out that Jeffrey conditioning is a logical consequence of the basic laws of probability and the constraint (sometimes called *rigidity*) that one's probabilities on a partition $\{e_i\}$

conditioning reduces to "strict" conditionalization as described above whenever $pr^*(e_i) = 1$ for exactly one element of the partition.

There are two basic strategies for justifying these various conditions as requirements on degree of belief: pragmatic and epistemic. The first takes degree of belief to be essentially constituted by its links to action and decision. The second does not.

17.1 Justifying Subjective Probabilities—Pragmatic Approaches

Early subjectivists such as Ramsey (1926/1990) and de Finetti (1937/1964) understood degree of belief in pragmatic terms; as Ramsey put it, "The degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it." A simple and common way of understanding this is in terms of betting dispositions. For example, let us call a bet that pays \$1 if p is true, and nothing otherwise, a "unit bet" on p. We will suppose that you can set a "fair" price for such a bet, i.e., a price at which you would be indifferent between buying and selling.[5] If \$$x$ is that price, then on the betting dispositions view, x would be your degree of belief in p. Now, it can be shown that if your degrees of belief so defined do not obey the laws of probability, you will accept bets as "fair" that result in certain loss for you. For example, if $x < 0$, then you will regard it as "fair" to give someone $-\$x > 0$ to take a bet in which you pay that person \$1 if p is true and nothing otherwise. Thus, you will lose at least $-\$x > 0$ for sure. If $x > 1$ for some logically true p, then you will pay more than \$1 for a bet that will only pay you \$1, thus losing $\$(x - 1) > 0$ for sure. If $x < 1$ for some logically true p, then you will sell a bet on p for \$$x$ and then will have to pay out \$1, thus losing $\$(1 - x) > 0$ for sure. Finally, let p and q be two logically exclusive statements, where \$$x$ is your fair price for a unit bet on p, \$$y$ your price for a unit bet on q, and \$$z$ your price for a unit bet on $(p \vee q)$. Suppose that $x + y > z$. Then you will pay $\$(x + y)$ for two unit bets, one on p and the other on q, and sell a unit bet on $(p \vee q)$ for \$$z$. Thus, your net payoff will be $\$[z - (x + y)] < 0$ for sure, as the following table shows.

p	q	Bet on p (bought for \$$x$)	Bet on q (bought for \$$y$)	Bet on $(p \vee q)$ (sold for \$$z$)	Total payoff
T	F	$\$1 - x$	$-\$y$	$\$z - 1$	$\$z - (x + y)$
F	T	$-\$x$	$\$1 - y$	$\$z - 1$	$\$z - (x + y)$
F	F	$-\$x$	$-\$y$	\$$z$	$\$z - (x + y)$

remain fixed. For it is a theorem of probability that when $\{e_i\}$ is some partition of the probability space, $pr^*(p) = \Sigma pr^*(e_i)pr^*(p|e_i)$. Therefore, if $pr^*(p|e_i) = pr(p|e_i)$ for all p and e_i, then $pr^*(p) = \Sigma pr^*(e_i)pr(p|e_i)$—which is Jeffrey's rule.

[5] Often it assumed that this applies no matter what the stakes S are; i.e., there's an x such that for any bet that pays \$$S$ if p and nothing otherwise, where S can be any real number not equal to 0, there's a subjectively fair price \$$xS$ for that bet, which determines one's degree of belief x in p.

If $x + y < z$, the same certain loss can be arranged by reversing who sells and who buys. Thus, the only way to avoid certain loss is if $x + y = z$.

This is what is commonly known as the "Dutch book" argument for the laws of probability. It assumes that (a) you can set fair prices for unit bets, and (b) these prices reflect your degrees of belief. If so, then if your degrees of belief do not obey the laws of probability, you will regard bets as fair that result in certain loss for you. Such degrees of belief are said to be "incoherent." (The equally important converse Dutch book theorem shows that if your degrees of belief obey the laws of probability, this can never happen, i.e., they are "coherent.") Thus, since fair betting prices must be non-negative, maximal for logical truths, and finitely additive, so (the argument goes) must degrees of belief. [6]

The literature on Dutch book arguments is quite extensive (see Hájek 2009 for a general review). There are Dutch book arguments for countable additivity (Williamson 1999), conditionalization (discussed in the next paragraph), Jeffrey conditioning (Armendt 1980; Skyrms 1987),[7] *reflection* (the thesis that one's current probabilities at t conditional on one's future ones at $t^\star$ should always be as follows: $pr_t(p|pr_{t*}(p) = x) = x$), and a "quasi"-Dutch book argument for *regularity* (the thesis, sometimes called *strict coherence*, that only logically true statements should get maximal probability).[8]

As noted earlier, conditionalization is a process in which one updates one's degrees of belief after learning something new for certain. Suppose you learn e is true (and that e is the logically strongest proposition you learn on that occasion), and that before learning e is true, your degrees of belief were defined so that $pr(h|e) = x$. Then, the argument goes, your new probability for h, $pr^\star(h)$, should be x. On the betting interpretation, $pr(h|e)$ is taken to be the fair price you'd assign to a "conditional" bet that pays \$1 if h, and nothing otherwise, but is "called off" (with the amount paid for the bet refunded) if e turns out to be false. The Dutch book argument for conditionalization (first stated by David Lewis, and reported in Teller 1973) is that if you have a rule for updating your degrees of belief that is not conditionalization, then even if your degrees of belief obey the laws of probability at all times (i.e., if for each time t there's a probability function that gives your degrees of belief at t), you will regard a set of bets as fair when they result in certain loss for you. In particular, suppose your rule says that you should update $pr(h)$ to $pr^\star(h)$ upon learning e for certain (and nothing more), and that $pr^\star(h) = y < pr(h|e) = x$, your current conditional probability for h given e. Then you would regard it as fair to buy a conditional unit bet on h given e for \$$x$ now and later (if you learn that e) will regard it as fair to sell a unit bet on h for \$$y$. Finally, if your probability for e is now z, then you'd regard it as fair to buy a bet that pays \$$(x - y) >$ 0 if e is true and nothing otherwise for \$$z(x - y)$. However, the total result of accepting all three bets would be certain loss, as the following table shows.

[6] The classic statements of the argument are Ramsey (1926/1990) and de Finetti (1937/1964). Shimony (1955) is also useful; the converse Dutch book argument is proven in Kemeny (1955) and Lehman (1955).

[7] Jeffrey's rule also has non-pragmatic defenses of the sort discussed in the next section. Diaconis and Zabell (1982) defend it as the unique rule that gives the "closest" function that meets the constraints of the new distribution over the e_i. Also notable is van Fraassen's (1986, 1989) symmetry argument for Jeffrey conditioning.

[8] Regularity and reflection are discussed elsewhere in this volume (Kotzen) (29). It is worth mentioning here, though, that regularity is lost as soon as one conditions on some contingent proposition e learned through experience (since after conditioning on e, $pr(e) = 1$).

h	e	Buy bet on h given e now for $\$x$	Buy bet on e now for $\$z(x - y)$	Sell bet on h then for $\$y$	Total
T	T	$\$1 - x$	$\$(1 - z)(x - y)$	$\$y - 1$	$-\$z(x - y)$
T	F	0	$-\$z(x - y)$	n.a.	$-\$z(x - y)$
F	T	$-\$x$	$\$(1 - z)(x - y)$	$\$y$	$-\$z(x - y)$
F	F	0	$-\$z(x - y)$	n.a.	$-\$z(x - y)$

Dutch book arguments have received a wide variety of criticisms. For example, one way to avoid Dutch books is never to bet, or at least avoid it if some shady character offers you a collection of bets that you notice collectively leads to sure loss. Some (e.g., Schick 1986) have argued against the "package principle" that if you consider each individual bet in a collection to be fair, you'll regard the whole collection (book) of bets as fair. It also makes no sense to make bets on propositions that can't be settled decisively, to everyone's agreement (e.g., God exists) or practically in one's lifetime (e.g., global warming will exceed 7°C in 2100). In the case of the conditionalization argument, you can avoid sure loss by not announcing to others your updating rule, or refusing to make the bet on h after learning e.[9] Hájek (2008) points out that incoherence (having degrees of belief that can result in your being on the losing side of a Dutch book) has a plus side, in that if you have such degrees of belief, you can also end up on the *winning* side of the Dutch book (in what he calls a "Czech book").

One response to all this is that it is best not to take Dutch book arguments too literally, and regard them instead as dramatizations of issues deeper than winning or losing money. For example, Ramsey (1926/1990) claimed that the flaw revealed by vulnerability to Dutch books is that your choices "depend on the precise form" in which the options are offered. More recently, Brian Skyrms (1984) has put the point as follows: you think of the bets as advantageous when described one way, but disadvantageous when described another way, when the second way is logically entailed by the first. For example, you might regard D_1 = "I buy a unit bet on p for 25¢, a unit bet on q for 50¢, and sell a unit bet on $(p \vee q)$ for 60¢" for two logically exclusive statements p and q as an equally advantageous (fair) arrangement for both you and your betting partner, but D_2 = "I lose 15¢ for sure" as a definite disadvantage for you. However, D_1 logically entails D_2. You regard certain loss as bad, when described as such, but you do not regard as bad a set of bets logically implying certain loss. A similar defense might be made for diachronic Dutch book arguments. Thus, it might be argued, it is no defense of a non-conditioning rule that one can avoid Dutch books by simply never making bets. Nor is it a defense that one can avoid a sure loss by not selling the bet on h later on, once e is learned for sure. The point is that one has adopted a rule which across time evaluates bets as individually advantageous or fair when collectively they *cannot* be, as a matter of logic.[10]

[9] Some representative publications that discuss these difficulties include Howson and Urbach (1989), Earman (1992), Maher (1993), and Hájek (2005). Levi (1987) and Maher (1993) critique the diachronic Dutch book.

[10] Worth mentioning is that van Fraassen (1989, 1990) has held that although it is true that if one adopts an updating rule, it must be conditioning (or its generalization, Jeffrey conditioning), it is

Whatever the force of such responses[11], the betting interpretation of degree of belief assumed so far (in particular the assumption that fair betting prices match degrees of belief) is in fact limited in application; it may be approximately true for some people in certain circumstances (e.g., for those not averse to betting, when the stakes are small, etc.), but it is not generally applicable. It does not take into account the subjective value of money (which varies from person to person, and with the amount—e.g., money has "diminishing marginal utility" so that each additional dollar gained is worth less than the preceding one), nor does it take into account the value or disvalue placed on betting itself, etc. Thus, a move to *expected utility theory* is necessary.

Accordingly, many subjectivists who follow Ramsey in taking degree of belief to be a pragmatic property, related essentially to how we are disposed to act, take degree of belief to be defined in terms of utility theory; in this framework, subjective probability and subjective value are intertwined. For example, Ramsey described a procedure for constructing a utility scale that keeps the general gambling framework but drops the assumption that money is a measure of value or that all bets are monetary ones. The basic idea can be illustrated simply. Suppose that A and B are options such that A is your highest preferred of all options and B your least; now let H be an "ethically neutral" event (an event for which you are indifferent between its occurring or not occurring, taken just by itself) that you regard as equally likely as not (e.g., a coin coming up heads, which would be "ethically neutral" if you have no intrinsic preference for heads over tails, or vice versa). We can take it that you think of H and ~H as equally likely if you are indifferent between the gambles (X if H, Y if not) and (Y if H, X if not) for all X and Y between which you are *not* indifferent (i.e., you strictly prefer one of them over the other). Then set the utility of A to 1 and the utility of B to 0, and set C to utility ½ if you are indifferent between C and the bet (A if H, B if not). Proceeding in that manner would allow the construction of a utility scale as fine-grained as you like. Once you have the utility u(E) of all events, we can define your probability for a statement p as follows: suppose you are indifferent between the options (X for certain) and (Y if p, Z if not); then $pr(p) = [u(\mathrm{X}) - u(\mathrm{Z})]/[u(\mathrm{Y}) - u(\mathrm{Z})]$. As Ramsey says, this definition "amounts roughly to defining the degree of belief in p by the odds at which the subject would bet on p, the bet being conducted in terms of differences of value as defined." Von Neumann and Morgenstern (1944) further developed this concept of cardinal (measurable) utility, using an explicitly axiomatic approach that set the standard for future developments.

In *The Foundations of Statistics*, Leonard Savage (1954) generalized such ideas in a rich framework that used the totality of one's preferences to define both subjective probability and utility. Savage proposed seven basic axioms of preference, and showed that any preference ordering that met those conditions could be represented by a unique probability function and utility function (the latter "unique" up to arbitrary positive linear transformations) combined according to the paradigm of mathematical expectation. So, if f is an "act," identified with a function from a state-space $\{s_i\}$ to an outcome-space $\{o_i\}$, the expected utility of f, which we will notate EU(f), is defined as $\Sigma pr(s_i)u(o_i)$. Intuitively, the s_i are various ways the world could turn out to be, and each o_i is the outcome if s_i obtains. Savage's representation theorem was thus as follows:

rationally acceptable sometimes not to follow any updating rule at all. The key point is that the diachronic Dutch bookie can only ensure gain if he knows your (incoherent) updating rule *in advance*.

[11] Hájek (2008) makes the point that absent a proof that obeying the laws of probability makes you invulnerable to such "inconsistencies," responses of this sort are indecisive.

> If the basic axioms are satisfied by a person S's preferences, then there is a unique probability function *pr* on $\{s_i\}$ and a utility function *u* (unique up to a positive linear transformation) on $\{o_i\}$ defining a utility function EU such that for any acts *f* and *g*, S prefers *f* to *g* if and only if EU(*f*) > EU(*g*).
>
> (Savage 1954: pp. 79–80)

(For the full background to the theorem, see Savage 1954 chapters 1–5.) On this approach, we can take *pr* to be S's subjective probabilities and *u* to be S's subjective values. This is the descriptive part of the project (i.e., defining what a person's degrees of belief and desire are). The normative part lies in taking the basic axioms to be requirements of rationality. For example, one's (strict) preferences should be asymmetric and transitive. One further requires that acts be capable of being placed into a linear ordering, and that preferences be rich enough so that a unique correspondence can be set up with the real numbers. A particularly notable feature of Savage's system is his independence axiom, motivated by the "sure thing" principle, which says that a person's preferences between acts *f* and *g* should be determined solely by the states at which their outcomes differ. Thus, let W be the set of states to which *f* and *g* assign the same outcome. Now let $f^\star$ and $g^\star$ be exactly like *f* and *g* outside of W, and let $f^\star$ and $g^\star$ agree with each other inside W, though *f* may not always agree with $f^\star$ inside W (and the same for *g* and $g^\star$). Then Savage requires that S prefers *f* to *g* if and only if S prefers $f^\star$ to $g^\star$. The following is an example of this (where the numbers indicate the utility of the outcomes, and $W = \{s_3\}$).

	s_1	s_2	s_3
f	5	10	3
g	3	15	3
$f^\star$	5	10	20
$g^\star$	3	15	20

Thus, one can prefer *f* to *g* and $f^\star$ to $g^\star$, or vice versa, but one cannot prefer *f* to *g* and $g^\star$ to $f^\star$.

Since Savage's achievement, which extended Ramsey's approach, there have been a large number of preference theories and associated representation theorems. Among the most notable is Richard Jeffrey's *The Logic of Decision* (1965), which does away with Savage's separate state-spaces and outcome-spaces in favor of a unified set of propositions, with acts simply being associated with particular propositions in that set. The result is that probabilities and utilities are even more tightly related than in Savage; under certain conditions, probabilities are not uniquely determined by preferences and are "unique" only up to fractional linear transformations defined partially in terms of utilities.

Thinking of subjective probabilities as being defined by preferences and utilities via representation theorems raises several issues. One is how to identify a person's degrees of belief, when the person's preferences do *not* obey the axioms of a given preference theory. Another is whether a particular set of axioms is really rationally required. Finally, even if a person obeys the axioms, and so can be *represented as* having a certain subjective probability

function, the question remains whether that means the person *really has* those subjective probabilities. Let us consider these questions in turn.

Soon after Savage's pioneering work was published, Maurice Allais (1952/1979) gave an example of a set of preferences that violated Savage's axioms, but seemed "reasonable" in some sense. Certainly, many people have the preferences he described, as subsequent research has shown. Consider choosing between A and B, and C and D.

Option A: You get $1,000,000 for sure.
Option B: You have an 89% chance of getting $1,000,000, a 10% chance of winning $5,000,000, and a 1% chance of getting nothing at all.
Option C: You have an 11% chance of winning $1,000,000; otherwise, you win nothing.
Option D: You have a 10% chance of winning $5,000,000; otherwise, you win nothing.

Many people prefer A to B, and D to C. However, note that we can represent these as follows, letting $pr(s_1) = 10\%$, $pr(s_2) = 1\%$, and $pr(s_3) = 89\%$.

	s_1	s_2	s_3
A	$1,000,000	$1,000,000	$1,000,000
B	$5,000,000	$0	$1,000,000
C	$1,000,000	$1,000,000	$0
D	$5,000,000	$0	$0

If Savage's axioms hold, then one can prefer A to B and C to D, or vice versa, but one cannot prefer A to B and D to C. It follows then that anyone who has the latter preferences (a common situation) cannot be represented in Savage's system. (Note that nothing need be assumed about the utility curve for monetary gain.) Does this mean such a person has no degrees of belief? Or that they have degrees of belief, but these do not obey the laws of probability? If so, what determines what their degrees of belief are? Finally, is it really irrational to have such preferences? (To many, such as Allais, such preferences have not seemed irrational.)

The problem has grown more acute over the years as empirical research (see e.g., Kahneman, Slovic, and Tversky 1982) has indicated the extent to which people violate the various laws of probability, and researchers have questioned the response that such violations are always irrational. Kahneman and Tversky's (1979) own "prospect theory" is one proposed alternative to expected utility, but a large number of other such approaches exist (see Fishburn 1988 for an extensive survey). To cite just one example, Fishburn's SSB (skew-symmetric bilinear) theory combines subjective probabilities with utilities so as to allow cyclical preferences. Here's a brief account of how it works. In this theory, acts are represented by a set P of probability distributions over outcomes. (The probability distributions are true probabilities, i.e., they obey all the laws of probability.) The strict preference relation $\succ$ over these distributions is assumed to be continuous (C1), convex (C2), and symmetric (C3), in the following senses.

C1. If p is strictly preferred over q, and q over r, then there is a mixture m of p and r with weights a and $(1 - a)$, respectively, such that m is indifferent to q.

C2. (a) If p is preferred-or-indifferent to both q and r, and strictly preferred to at least one of them, then p is strictly preferred to any mixture of q and r.
(b) If p is indifferent to both q and r, then p is indifferent to any mixture of q and r.
(c) If q and r are preferred-or-indifferent to p, and at least one of them is strictly preferred to p, then any mixture of q and r is strictly preferred to p.

C3. If p is strictly preferred to q, q to r, and p to r, and q is indifferent to a 50:50 mixture of p and r, then
(a) any mixture of p and r, with weights a and $(1 - a)$, respectively, is indifferent to a 50:50 mixture of p and q iff
(b) any mixture of r and p, with weights a and $(1 - a)$, respectively, is indifferent to a 50:50 mixture of r and q.

Fishburn (1988) shows that any preference relation $\succ$ that satisfies C1–C3 can be represented by a skew-symmetric, bilinear (SSB) functional f on P × P such that $p \succ q$ iff $f(p, q) > 0$, and that this is unique up to multiplication by a positive constant. He gives the following example of a preference cycle that SSB theory permits. Suppose that someone is faced with the following options.

A: A 6% raise with 90% probability, no raise otherwise
B. An 8% raise with 72% probability, no raise otherwise
C. A 10% raise with 61% probability, no raise otherwise

A person might prefer A to B because there's only a small (2%) increase in the raise from A to B with a big drop in probability of getting it, and for similar reasons she might prefer B to C. However, when comparing A to C, C looks better because it represents a significant improvement in the raise (4%), which represents a kind of "threshold" for her, resulting in a significant and noticeable increase in her standard of living that more than compensates for the lower probability of getting it. Fishburn (1988: pp. 74–5) gives as an example of an SSB function $f(x, y)$ that represents these preferences $v(x - y)$, where the values given by v to 2, 4, 6, 8, and 10 are 1.2, 3, 6, 8, and 9.8, respectively. Thus, v(A,B) = 0.1584, v(B,C) = 0.04552, and v(C,A) = 0.1388, all of which are positive.[12]

It cannot be said that such preferences are "inconsistent" in any literal, logical sense. For example, the person who has cyclical preferences and conforms to Fishburn's axioms does not give positive credence to any contradiction. (Note that the degrees of belief here are coherent probability functions.) Now, it is true that cyclical preferences violate the axioms of expected utility theory. Cyclical preferences have been widely regarded as irrational on both intuitive and pragmatic grounds. Intuitively, preferring A to B to C to A seems repugnant,

[12] Here's how this works for v(A,B). Let a and b designate the raises associated with A and B, respectively (i.e., 6, 8). In either case the alternative is no raise, so $a' = b' = 0$, with $v(0) = 0$. The difference $a - b$ is -2, so $v(a,b) = v(a - b) = v(-2)$, which given skew symmetry is $-v(2) = -1.2$. We have set it so that $v(a) = 6$ and $v(b) = 8$. Finally, let $p = 0.9$ and $p' = .72$ (the probabilities associated with the raises for A and B). Now v(A,B) is bilinear in both arguments. Considering just A for the moment, this implies v(A,B) = $v((pa + (1 - p)a'), B) = pv(a,B) + (1 - p)v(a',B)$; and, since $B = p'b + (1 - p')b'$, bilinearity implies further that $v(A,B) = p[p'v(a,b) + (1 - p')v(a,b')] + (1 - p)[p'v(a',b) + (1 - p')v(a',b')]$. Thus, v(A,B) = (.9((.72(-1.2))+(1–.72)6))+((1–.9)(.72(-8)+(1–.72)0)) = 0.1584. The other terms v(B,C) and v(C,A) are calculated analogously.

and confused. Pragmatically, it is sometimes claimed that a person with such preferences would be a "money pump," i.e., that the person, starting with A, should be willing to pay a small price \$$c$ to obtain C instead, and then should be willing to pay a small price \$$b$ to obtain B, and then a small price \$$a$ to obtain A …putting himself back where he started but \$$(a + b + c) > 0$ poorer. And then the cycle can begin again (assuming his preferences remain the same).

However, along the lines of Schick's (1986) rejection of Dutch book and money pump arguments, it might be argued that such a person might simply see that he'll end up where he started at some point, and refuse to continue. If at every point such a person gains by his own lights, but he stops before any cycle is completed, how is he irrational? Horowitz (2006: p. 107) has argued that Fishburn's theory provides some grounds for thinking cyclical preferences are *not* necessarily irrational: "The argument that cyclic preferences are rationally acceptable rests, ultimately, on the availability of a utility theory that seems quite reasonable [i.e., Fishburn's SSB theory], but that tolerates preference cycles."

Now, I am not arguing that cyclical preferences are in fact reasonable (I would deny this[13]), but, it must be admitted that such preferences *exist*. Hence, it could be the case that a person does not conform to the basic axioms of expected utility theory but *does* conform to Fishburn's SSB axioms. So, such a person could have subjective probabilities (i.e., coherent degrees of belief), *even though they have preferences that are wildly out of conformance with the axioms of expected utility theory*. The lesson here is as follows: there are a large number of alternatives to expected utility. Some of these incorporate subjective probabilities, combining with utilities in ways different from expectation (e.g., Fishburn's SSB theory). Others (which I have not dealt with here) incorporate sub-additive "probabilities." Each has its own set of axioms and representation theorems. So, it follows that subjectivists who take the pragmatic approach and want to argue for probabilism based on representation theorems must defend particular sets of preference axioms as reasonable, for normative purposes, and perhaps other axiom-sets as descriptively adequate, to specify people's degrees of belief when they fail to have "reasonable" preferences.

There is one last problem to note. Suppose that some set of axioms for expected utility can be defended. An argument for probabilism that is based on the representation theorems of expected utility theory is valid only if it can be shown that having subjective probabilities is needed to avoid violations of these axioms. However, Zynda (2000) shows that this is not so. In his example, non-additive degrees of belief are shown to produce preferences that agree with those of expected utility theory. The trick is to combine these with utilities in an unorthodox manner. In particular, preferences between acts (conceived of as maps from a state-space $\{s_i\}$ to an outcome-space $\{o_i\}$, as in Savage's system) are defined by "valuation," not expected utility, where b is a sub-additive belief function with values ranging from 1 to 10 and the valuation V(A) of an act A is defined as $\Sigma_i(b(s_i)u(o_i) - u(o_i))$. So defined, it follows that EU(A) > EU(B) iff V(A) > V(B). Eriksson and Hájek (2007) report a suggestion along these lines by Joyce, namely, that where $\mathrm{EU}(A_i) = \Sigma_j\, u(s_j \,\&\, A_i)pr(s_j)$ is the expected utility of act A_i, one can always define a non-expected valuation ("schmexpected utility") $\mathrm{V}(A_i) = \Sigma_j\, [f(A_i, s_j)u(s_j \,\&\, A_i) \times pr(s_j)/f(A_i, s_j)]$, where "$f(A_i, s_j)$ can be any non-zero function you like," that reproduces the preference ordering of EU. (Here, the

[13] Briefly, if a person refuses to trade C for A when they prefer A to C because they'll lose as the result of some sequence of prior exchanges, in what sense do they *now* prefer A to C? Preferring A to C *just means* that you regard it as advantageous to obtain A over C.

term $pr(s_j)/f(A_i, s_j)$ defines your degrees of belief $b(s_j)$.) Thus, conforming to the axioms of expected utility theory doesn't establish you have subjective probabilities, without some argument eliminating all varieties of "schmexpected utility" as admissible representations of your beliefs and values.

17.2 Justifying Subjective Probability: Non-pragmatic (Epistemic) Approaches

It is possible to argue against pragmatic interpretations of degree of belief on various grounds. For example, one might argue that many contexts in which we make probabilistic judgments (such as the initial plausibility of a scientific theory, or the degree to which evidence would increase our confidence in a theory) do not evidently involve (even implicitly) considerations of betting or decision-making or preference rankings, as outlined in the previous section, and that, moreover, such considerations are not needed to make these judgments. Indeed, the complexities that became apparent in the previous section (e.g., with respect to the status of the representation theorems, and the various alternatives to expected utility) might lead one toward the view that decision-theoretic issues are so many red herrings, if our primary aim is to provide grounds for evaluating the *epistemic* status of our degrees of belief. What, then, are the alternatives to the pragmatic view that, in Ramsey's words, "The degree of a belief is a causal property of it, which we can express vaguely as the extent to which we are prepared to act on it?" As was the case before, there are both descriptive and normative issues to address. The *descriptive* problem is giving an account of what our degrees of belief are, and how we know what they are, if they are not essentially connected with action, or dispositions to act. There are several possibilities. A non-pragmatist might, for example, take "the degree of a belief" to be a "felt" quality (beliefs come in different "intensities," of which we are aware by introspection), or the result of a conscious mental act (e.g., making an epistemic judgment), or as a primitive psychological state that we "express" via such judgments (and in other ways, too). A defender of the pragmatic approach might respond that we often don't know how confident we are until we are forced to act, or to think seriously about how we might act. Certainly, it might be argued, self-deception is possible with degrees of confidence just as it is with belief and desire generally. The non-pragmatist might concede that we are fallible with respect to knowing our degrees of belief, while maintaining that they are only *causally* related to other, action-oriented mental states (such as preference and degrees of desire), and are not *constituted* by such relations. In any case, while these descriptive issues are important, we will leave them behind in what follows[14] (assuming that some non-pragmatic descriptive account of degrees of belief can be given) and concentrate on the *normative* problem, namely, giving a non-pragmatic *justification* for probabilism.

[14] Eriksson and Hájek (2007) give a good survey of the various (pragmatic and non-pragmatic) answers to the ontological question regarding degree of belief. They themselves endorse the primitivist approach.

One thing we want of our beliefs is that they be *true*; by analogy, then, what we would want of our degrees of belief is that they be more or less *accurate*, or *close* to the truth, in some sense. For example, even though subjectivists don't think probabilities are long-run limiting relative frequencies, they do hold that observed frequencies should affect our degrees of belief, and that our degrees of belief should (at least potentially) "match" observed frequencies somehow. The simplest development of this idea is the *calibration* argument (see, e.g., van Fraassen 1983). Suppose that a weatherman judges that the probability of rain is x% on days when conditions C obtain; then he is perfectly calibrated with respect to that judgment if it in fact rains on x% of the days on which conditions C obtain. Perfect calibration is of course not to be expected; by analogy, one does not expect a fair coin always to yield ½ heads and ½ tails. (Indeed, the probability of getting exactly 50% heads is zero when the number of tosses is odd, and for an even number of tosses, actually decreases as the number of tosses increases, and converges to zero.) Nonetheless, a "close" fit is desirable, closer as the sample size increases, and (here is the crux of the argument) it should be at least logically *possible* that one's probability judgments be a perfect fit to the observed frequencies. The calibration argument, then, is that if one's probability judgments do not conform to the laws of probability, then they cannot (as a matter of logic) be a perfect fit to the observed frequencies. That is because relative frequencies are necessarily non-negative, maximal for logical truths, and additive. So, for example, if one considers a variable *cloud cover* with possible values *sunny*, *partly cloudy*, *mostly cloudy*, and *cloudy* (represented as *S*, *P*, *M*, and *C*, respectively), then the relative frequency of each of these cannot be less than 0%; the number of days on which at least one of these values obtains must be 100%; and the relative frequency of *partly-or-mostly-cloudy days* ($P \cup M$) must be the sum of the relative frequencies of the partly cloudy days (P) and the mostly cloudy days (M). So, if our degrees of belief do not have these properties, it follows that they cannot match the observed frequencies, no matter what these might be.

Now, calibration arguments such as the one just given are limited in scope; e.g., they only apply when relative frequencies make sense, i.e., when one has a repeatable event-type such as coin tosses, rainy or cloudy days, etc. They don't (at least obviously) make sense when one is speaking of scientific theories. (Einstein's general theory of relativity isn't true on some days and false on others.) Moreover, fit to observed frequencies (which as a matter of fact are always rational numbers, since our observations are always finite) would seem to preclude our having a degree of belief of, say, $1/\pi$. This seems highly undesirable, particularly when quantum mechanics sometimes gives us probabilities that are irrational numbers. So, what is needed is some measure of accuracy that is broader in scope than calibration, and some proof that one will always do better (in terms of accuracy, so measured), no matter how the world might be, if one's degrees of belief obey the laws of probability than if they do not.

Building on earlier work by Brier (1950), Lindley (1982), and others (right back to de Finetti), Joyce has recently developed a non-pragmatic argument for probabilism of this sort, i.e., one based on a notion of *gradational accuracy* conceived in terms of closeness to *truth*, rather than of closeness to observed relative frequencies. The goal is to show that (a) for any incoherent degree-of-belief function, there's a coherent one that is at least as accurate in all possible worlds, and more accurate in some, and (b) no coherent belief function is dominated by an incoherent one in this way. The first version of his argument (Joyce 1998) assumed several intuitively motivated constraints (e.g., regarding structure, extensionality, normality, dominance, weak convexity, and symmetry), some of which were

criticized by Maher (2002) on the grounds that they were not sufficiently well motivated, and that moreover the argument did not prove its intended result. In response, Joyce (2009) revised his argument considerably. It is that revision that will be summarized here.

First, it's assumed that a belief function b is *inadmissible* if there's a belief function b^* that is at least as accurate as b in every possible world, and more accurate than b in at least one possible world. Secondly, *gradational accuracy* of an estimate is defined so that an estimate *Est* is "uniformly closer" to the values of variables in $\{F_i\}$ than another estimate *Est** if *Est*(F_i) is always at least as close to the true value of F_i for all F_i, and sometimes closer (i.e., strictly closer for at least one F_i), than *Est*$^*(F_i)$. Thirdly, it's assumed that accuracy for belief functions involves estimation of *truth* (so that any plausible measure of accuracy will be "truth-directed").[15] For example, if $b(p) = 0.5$ and $b^*(p) = 0.6$ for some proposition p true at world w, b^* is "more accurate" with respect to p at w (letting *true* $= 1$).[16] If b^* is at least as accurate as b for all p at w, and more accurate for at least one p at w, then b^* is "uniformly closer to the truth" at w. Finally, Joyce assumes *coherent admissibility*, which says that no rationally defensible epistemic scoring rule (i.e., specific measure of accuracy) can judge a coherent degree-of-belief function b to be inadmissible (the idea being that one shouldn't, in Joyce's words, make "coherent credences irrational *a priori*"). From this, it can be shown that any incoherent belief function b is inadmissible—in other words, for each incoherent belief function, there's a coherent one that's at least as accurate at every possible world and strictly more accurate at some, with no incoherent belief function being more accurate than a coherent one in this way.

Joyce's argument depends on a relatively minimal set of assumptions; nonetheless, one might question its reliance on the assumption that all probability functions are admissible. Why should we simply assume this? Hájek (2008) makes the important point that adopting coherent admissibility as a premise amounts to simply assuming the converse theorem to Joyce's (i.e., that if you have coherent degrees of belief, they have gradational accuracy). Leitgeb and Pettigrew (2010a, 2010b) also criticize this assumption of coherent admissibility, and seek to prove a result similar to Joyce's without it. They show that, even if one uses an incoherent degree-of-belief function b to estimate inaccuracy, probability functions will still do a better job at minimizing *expected* inaccuracy, as defined by b. First, they define local inaccuracy (distance-from-truth with respect to a particular proposition p at world w) and global inaccuracy (overall distance-from-truth of a belief function at world w), and then focus on the idea of *expected* inaccuracy of both sorts relative to a belief function b. Thus, where $I(p,w,x)$ is the local inaccuracy of degree of belief x in p at w, the expected inaccuracy of that degree of belief x in p according to b will be $\Sigma b(w)I(p,w,x)$, where the sum is calculated over all epistemically possible worlds (i.e., the set of all worlds for which $b(p) > 0$). Moreover, one can use one's belief function b to calculate the expected (global) inaccuracy of another function b^*. Thus, where $G(w,b^*)$ is the global inaccuracy of b^* at w, the expected inaccuracy of b^* according to b will be $\Sigma b(w)G(w,b^*)$. Leitgeb and Pettigrew argue for certain constraints on allowable inaccuracy measures; e.g., they argue

[15] As Joyce (2009: p. 285) notes, calibration violates truth-directedness, since "my credences might be uniformly closer to the truth than yours, and you might still be better calibrated to the frequencies than I am."

[16] A "possible world" is modeled as a vector that assigns 1 or 0 to each proposition in a partition $\{p_i\}$ such that each proposition in the space is a union of one or more elements of this partition.

that the only legitimate local measures $I(p,w,x)$ are those that have the form $\lambda(\chi_p(w) - x)^2$, where $\chi_p(w)$ is the characteristic function that takes the value 1 or 0 depending on whether p is true or false at w, and λ is any positive real number. (A similar constraint is argued to hold for the global measure $G(w,b^*)$.) Finally, they assume that one should *minimize the expected inaccuracy* (in both senses) of one's degrees of belief at time t, as calculated according to one's belief function at t. Intuitively, I should not assign a degree of belief x to any proposition p (i.e., set $b(p) = x$) if according to b, another degree of belief x^* would be a better estimate of p's truth (*synchronic expected local accuracy*). Also, I should not have a belief function b if according to b, another function b^* is overall a more accurate estimate of the truth (*synchronic expected global accuracy*). (The former assumption logically entails the second.) What Leitgeb and Pettigrew prove is that synchronic expected local accuracy implies that one's degrees of belief ought to be probabilities. In other words, if one's degrees of belief are incoherent, then one will judge (using those degrees of belief) that a coherent belief function would do a better job at minimizing expected inaccuracy.[17]

17.3 Coherence, "Rationality," and Evidential Support

We discussed in previous sections arguments that the various constraints of probabilism are requirements of rationality. In this section, we will consider the question: What does it mean to say that coherent degrees of belief are "rational?" A puzzle arises because the constraints probability theory puts on degrees of belief apparently require a kind of *logical omniscience*. One must give maximal probability to all logical truths, which seems to require that one knows what they are.[18] Moreover, a theorem that follows from the basic laws of probability states that if p logically implies that q, then $pr(p) \leq pr(q)$, which further implies that if p and q are logically equivalent, $pr(p) = pr(q)$. All this seems to rule out logical (or mathematical) learning, in which one may initially fail to recognize logically equivalent statements or logical truths as such, and so gives different degrees of credence to equivalent statements, or less than full credence to logical truths. Surely, however, one can be "rational" without being logically omniscient, and logical and mathematical learning can be as rational as any other learning. For example, mathematicians were highly confident Fermat's last theorem was true before Andrew Wiles proved it in 1995, but it would seem reasonable for them before this to have allowed that it might be false. And certainly a non-mathematician might reasonably harbor some doubt. But there is no possible world in which Fermat's last theorem is false.

This issue becomes important in the well-known *problem of old evidence*. Here is the background. Typically, subjectivists assume that evidence e supports h for an agent S at t if $pr_t(h|e) > pr_t(h)$ (where pr is S's probability function at t). As noted above, subjective probabilities vary from person to person, and they also change over time; thus,

[17] Leitgeb and Pettigrew also prove some interesting diachronic results that unfortunately there is insufficient room to discuss here. E.g., they question the scope of Jeffery conditioning, and propose an alternative rule for when the rigidity condition (discussed earlier) does not hold.

[18] Or at least, whenever one does not know, that one guesses and always hits it right by sheer luck.

it seems that evidential support must as well. In particular, if an agent always updates by conditionalization, and sets $pr(e)$ to 1 upon learning at t that evidence e obtains, then $pr_{t*}(h|e) = pr_{t*}(h)$ for all times $t^* \geq t$. Thus, if we take evidential support to obtain exactly when $pr(h|e) > pr(h)$, it follows that evidence no longer supports a hypothesis once it is "known."[19] This seems to go against our ordinary way of speaking about evidential support. The cosmic background radiation didn't only support the big bang theory before this radiation was discovered; it still does! Now, there are two problems here, arising from the fact that "support" has two senses: one is to define a *relation* of evidential support between statements; the other is to explain an *event*, i.e., how evidence makes (or made) scientists more confident in their theories.[20] Regarding the latter problem, it is plausible for a subjectivist to say that when scientists gain some new evidence, it can make them more confident in a theory, but that simply reconsidering the same evidence again after that won't make them any *more* confident. In this way, evidence no longer "supports" a theory (in the sense of further increasing scientists' confidence in it) after it is known. The former problem has its own difficulties for the subjectivist. Those who believe in logical probability can at least try to specify an initial evidence-free situation by a unique "ur-prior" pr_0 (or a well-defined, restricted class of such functions) and say that the conditional probabilities $pr_0(h|e)$ give the degree to which e objectively supports or confirms h. Subjectivists however cannot appeal to a single such function, since they allow that "ur-priors" can differ from person to person. Indeed, the most liberal subjectivists (who regard coherence as the only constraint on rational degrees of belief) would allow that any probability function can be a rational "ur-prior." So, even if they can get around the relativity of evidential support to time, they cannot get around its relativity to persons.

The problem of old evidence (first posed by Clark Glymour in 1980) specifically arises when evidence is "known" (where this is understood as "having probability 1") *before* a theory is even formulated. It seems that for subjectivists, such evidence cannot support a new theory, in either sense of "support" described above, since $pr(e) = 1$ before the theory was formulated, and thus $pr(h|e) = pr(h)$ from day one. Thus, Einstein's general theory of relativity cannot have been supported by the anomalous advance in Mercury's perihelion, since this had been known for decades before Einstein worked on his theory (and was known to Einstein as well),[21] and Darwin's theory of "descent with modification" through natural selection cannot have been supported by the many already known facts he put forward in *The Origin of Species* (e.g., about homologies, rudimentary organs, fossils, the geographical distribution of species and varieties, etc.) as support for his theory. There are two basic ways subjectivists have tried to solve this problem, with an important distinction existing within the second.

1. **Counterfactual Priors**: Say that e supports h for S at t if $pr^*(h|e) > pr^*(h)$ for the probability function pr^* that S would have had if S had not known that e at t (all else being equal, as much as possible)

[19] This way of speaking occurs commonly in the literature on the problem of old evidence but it is a bit misleading, since one can give probability 1 to propositions one does not know, and one can know propositions to which one gives high but not maximal probability.

[20] The difference between confirmation-relations and confirmation-events is explored more fully in Zynda (1995).

[21] See Roseveare (1982) for an interesting account of this.

2. **Non-Observational Learning**: Say that it is not the evidence *e* itself that supports *h* at *t*, but rather facts that are learned at *t* about the relationship between *h* and *e*, such as that *h* entails *e* or that *h* explains *e*. This can be implemented in two basic ways, either:

   a) $pr(h|h \text{ entails } e) > pr(h)$, with *S* then conditioning on "*h* entails *e*" in the normal way once this is learned, with the same going *mutatis mutandis* for *h*'s explaining *e*); or
   b) $pr^*(h) > pr(h)$ where pr^* is some probability function that results from *pr* and some rule other than conditioning for incorporating the information that *h* entails (or explains) *e*.

Theorists that support the first (counterfactual priors) solution include Howson and Urbach (1989); a similar suggestion due to Skyrms is reported by Eells (1990). Now, there do seem to be situations where one can come up with a definite answer to questions about one's counterfactual priors. The problem is that there are many situations where such priors may be impossible to determine, or might give the wrong results. If Darwin hadn't known the various facts about fossils, embryology, classification, geological distribution, etc. that he cites in the *Origin*, he would have to have not been the erudite naturalist he was. Similarly, it is difficult to see how Einstein would have been the competent physicist he was without knowing about the anomalous advance in Mercury's perihelion. Perhaps what would have happened, if they hadn't known about these things, is that they never would have formulated their theories at all. (And so, they could not have had these theories in their ur-prior's probability space.) This is a general problem since in science there are often longstanding unexplained phenomena, which everyone in a field knows about, and which new theories gain credence by (finally) explaining. Indeed, as Kuhn (1970) has pointed out, such "anomalies" often drive the search for new theories. In addition to these difficulties, there is another: *counterfactual* priors cannot explain how scientists *actually* become more confident in *h* based on old evidence *e*, since one's actual epistemic history depends on one's *actual* priors. So, we cannot explain how and why Einstein became confident he was on the right track when he discovered his general theory implied the anomalous advance in Mercury's perihelion by appealing to a counterfactual situation in which he did not even know about that anomalous advance.

The second approach (2a and 2b), which allows for the learning of "new" truths involving the relationship between *h* and old evidence *e*, avoids these problems, but runs into the difficulty of logical omniscience mentioned above. If *h* entails *e* deductively or mathematically, it seems that "*h* entails *e*" is a logical or mathematical truth, and hence it ought to be the case that $pr(h \text{ entails } e) = 1$, supposing statements about logical relationships such as "*h* entails *e*" are in the probability space at all (as solutions of the form 2a require). Even if they are not in the probability space (as solutions of type 2b allow), if *h* entails *e*, then the probability calculus requires that $pr(e|h) = 1$, which means that $pr(h|e) = pr(h)/pr(e)$, and, since $pr(e) = 1$ (since *e* is "known"), it follows that $pr(h|e) = pr(h)$. Therefore, to solve the problem in the second way (2a or 2b), one has to allow for forms of learning in which logical or mathematical truths initially receive non-maximal probability. Thus, to solve the problem of old evidence along the second route, one has to allow for incoherent degree-of-belief functions that do not fully obey the laws of probability.

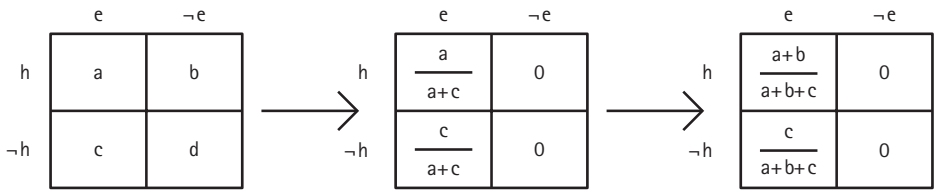


FIGURE 17.1 Applying the rule of reparation after conditioning on *e*.

Garber (1983; see also Jeffrey 1983/1992) originally proposed a solution to this problem in which the agent learns propositions of the form *h* entails *e* by ordinary conditioning on them; he showed that if we include such propositions in an enlarged probability space, in which "extrasystematic" statements of the form "*h* entails *e*" are treated as "atomic," then under certain conditions, $pr(h|h \text{ entails } e) > pr(h)$. The conditions in question allow that $pr(h \text{ entails } e) < 1$ (since they require only that truth-functional tautologies compounded from the "atomic" sentences get maximal probability) and require only that modus ponens is respected (e.g., they imply that $pr(e|h \,\&\, h \text{ entails } e) = 1$ whenever $pr(h \,\&\, h \text{ entails } e) \neq 0$.)[22] There was some controversy over the "extrasystematic" status of the statements "*h* entails *e*" and the liberality of Garber's conditions (for discussion, see van Fraassen 1988; Eells 1990, Earman 1992, and Zynda 1995). For example, the simple statement "$\forall xFx$ entails $\exists xFx$" can get any probability whatsoever in Garber's system, but an agent still has to give all truth-functional truths probability 1, no matter how complex they might be.

Jeffrey (1991/1992) proposed a new solution of the form (2b), in which a rule called "reparation" allows for a probability function *pr* to be revised when it is learned that *h* entails *e*, without the proposition "*h* entails *e*" actually being in the space itself. Here's an example of how it works.

In Figure 17.1 above, the leftmost box represents an "ur-prior" pr_0 in which all Boolean combinations of *h* and *e* get some positive probability, including the logically impossible $h \,\&\, \sim e$. In the middle box, the agent conditions on *e*, so that $pr(\text{-}) = pr_0(\text{-}|e)$. Later, the agent discovers that *h* entails *e*, and "repairs" the mistake of not recognizing this earlier by altering the middle distribution *pr* to what would have resulted if the probability *b* originally assigned to the logically impossible $h \,\&\, \sim e$ had been included in the $h \,\&\, e$ cell, and he had then conditioned on *e*. In other words, the distribution in the rightmost box (call it pr^*) is exactly what would have resulted from ordinary conditioning on *e* from an *alternative* ur-prior $pr_0{}^*$ in which $pr_0{}^*(h \,\&\, e) = (a + b)$, $pr_0{}^*(h \,\&\, \sim e) = 0$, $pr_0{}^*(\sim h \,\&\, e) = c$, and $pr_0{}^*(\sim h \,\&\, \sim e) = d$. The idea behind the rule of reparation is that what would've been appropriately applied to the ur-prior pr_0 to produce $pr_0{}^*$ ought to be applied at any given point later when the same information (namely, that *h* entails *e*) is learned.

Jeffrey's reparation rule can be stated generally as follows: Upon learning that *h* entails *e*, multiply the odds on *h* given *e* by $1/pr_0(e|h)$, set the odds on *e* given *h* to infinity (i.e., set $pr(e|h)$ to 1), and renormalize, leaving everything else unchanged. Applying this to the leftmost box (i.e., reparation of the ur-prior pr_0 *before* learning that *e*) results in the

[22] In other words, except for the "modus ponens" condition, "*h* entails *e*" is treated as if it were a simple, unstructured sentence *p*.

alternative ur-prior $pr_0{}^*$ specified above, since the odds on h given e in the leftmost box are a:c, and so multiplying this by $1/\ pr_0(e|h) = (a + b)/a$ yields revised odds on h given e of $(a + b)$:c. Setting the odds on e given h to infinity then results in $pr_0{}^*$. Applying reparation to the middle box (i.e., reparation *after* learning that e) works as follows. Since there $pr(e) = 1$, the odds on h given e are simply the odds on h, namely a:c. So, after applying Jeffrey's rule, the new odds on h will be $(a + b)$:c. This gives the distribution pr^* at right. Now, it is obvious that $pr * (-) = pr_0 * (-|e)$. So, reparation after conditioning gives the same result as conditioning after reparation. Now, note that reparation after conditioning results in an increase in the probability of h, since $pr^*(h) = (a + b)/(a + b + c) > a/(a + c) = pr(h)$. Thus, learning that h entails e can increase the probability of h even when $pr(e) = 1$.

Jeffrey's reparation rule was developed more generally by Carl Wagner (1997, 1999, and 2001). Wagner extends the rule to cases where the relationship between h and e is probabilistic (i.e., h makes e likely, without entailing it), and where the number of evidentiary propositions is countable, rather than finite. So, for example, instead of learning that h logically implies e, one might learn (on the basis of theoretical explanatory considerations alone) that e should have probability u if h is true, and v if h is false. Call this "explanatory" information E. In applying this to old-evidence situations, in which the probabilistic ("explanatory") relationship between h and e is learned after observational learning (e.g., learning by ordinary Jeffrey conditioning), Wagner retains Jeffrey's idea of utilizing an ur-prior pr_0, and the idea that observation- and explanation-based learning should commute; he bases the latter on the *uniformity principle*, which says that the same information should always be incorporated in the same way, whenever it is learned. Thus, if we learn that E after some observation-based Jeffrey conditioning has occurred, this information should be incorporated in the same way as if we'd learned E before the Jeffrey conditioning occurred. Importantly, the "same way" needs to be specified in terms of Bayes factors, not probabilities, in order for commutation to occur.[23] If we consider the case of how to update an ur-prior pr_0, upon learning E (and nothing more), Wagner argues that one ought to adjust pr_0 to a new distribution $pr_0{}^*$ such that $pr_0{}^*(e|h) = u$ and $pr_0{}^*(e|{\sim}h) = v$, while leaving the probability of h unchanged. Here's a numerical example that does this, letting $u = ¾$ and $v = ½$.

Statement	pr_0	$pr_0{}^*$
h & e	.1	.225
h & ~e	.2	.075
~h & e	.3	.35
~h & ~e	.4	.35

However, this change, when stated in terms of Bayes factors and applied to a different probability function pr, would not necessarily leave the probability of h unchanged.

[23] One objection to Jeffrey conditioning (see, e.g., Levi 1991) has been that the order in which one incorporates new information can affect the final results, if specified in terms of the posterior probabilities of the $\{e_i\}$ on which one Jeffrey conditions. Jeffrey (1988: pp. 235–7) notes however that his conditioning rule is commutative when specified in terms of ratios (odds). (See Wagner 2002 for further discussion of this issue.)

Indeed, the explanatory information E can increase the probability of h. Here is the particular proposal Wagner puts forward. First, where s_1 and s_2 are two statements in the probability space of pr_1 and pr_2, define the Bayes factor $\beta_{pr1,pr2}(s_1{:}s_2)$ to be $[pr_1(s_1)/pr_1(s_2)]/[pr_2(s_1)/pr_2(s_2)]$. Secondly, simplifying our notation a bit, let p be the ur-prior, $p^\star$ the modified ur-prior that incorporates E, q some later probability function derived from p (perhaps after much observational learning has occurred), and $q^\star$ the probability function that results from applying the generalized reparation rule to q. Then the following is the Bayes identity behind Wagner's generalized reparation rule.

For all propositions s_1 and s_2, $\beta_{q*,q}(s_1{:}s_2) = \beta_{p*,p}(s_1{:}s_2)$.

Now, in the case Jeffrey considered (a special case of Wagner's rule in which u moves to 1 and v remains fixed), h became more likely whenever it was learned that h entails e. In the more general case Wagner considers, the probability of h does not always increase upon learning that h (probabilistically) explains e. Here's a case where it does.[24] The Bayes factors for the shift above are $\beta_{p*,p}(h \,\&\, e{:}h \,\&\, {\sim}e) = 6$, $\beta_{p*,p}(h \,\&\, {\sim}e{:}{\sim}h \,\&\, e) = 0.32$, $\beta_{p*,p}({\sim}h \,\&\, e{:}{\sim}h \,\&\, {\sim}e) = 1.33$, and $\beta_{p*,p}({\sim}h \,\&\, {\sim}e{:}h \,\&\, e) = 0.39$. Let q be the function that results from Jeffrey conditioning p on the partition $\{e, {\sim}e\}$ with $q(e) = 0.6$ and $q({\sim}e) = 0.4$. Applying these Bayes factors to q results in $q^\star$; note that $q^\star(h) = 0.338 > q(h) = 0.283$. Thus, the explanation-based change from q to $q^\star$ increases the probability of h (even though the same change did *not* increase the probability of h in the move from p to $p^\star$).

Statement	q	$q^\star$
h & e	.15	.295
h & ~e	.133	.044
~h & e	.45	.458
~h & ~e	.267	.204

Note, finally, that $q^\star$ results from $p^\star$ by Jeffrey conditioning with the same ratios that led from p to q, where the factor for e is 1.5 and ~e is 2/3. Thus, generalized reparation and Jeffrey conditioning commute (so long as the latter is specified by ratios).[25]

The need to incorporate logical and mathematical learning in the problem of old evidence highlights the need for a general descriptive account of incoherent degrees of belief. In particular, the idea that the probability calculus presents an "ideal" of rationality that we ought to approximate as best we can implies a need for some notion of comparative incoherence, so that we can speak meaningfully of "getting closer" to the ideal of coherence. Zynda (1996) presents a means of partially ordering incoherent degree-of-belief functions (with the degree-of-belief space allowed to be countable) in terms of the extent to which those functions' assignments are coherent. He shows that based on this ordering, one

[24] Wagner (1997, 1999, 2001) proves several theorems about when generalized reparation will increase the probability of h and when it will not.

[25] Note that $q(e)/p(e) = 1.5$ and $q({\sim}e)/p(e) = 2/3$. Multiplying $p^\star(e)$ and $p^\star({\sim}e)$ by these factors and renormalizing gives $q^\star(e) = 0.75$ and $q^\star({\sim}e) = 0.25$.

always improves one's opinion (i.e., moves closer to the ideal of coherence) if one corrects a local incoherence in one's degrees of belief, even while other incoherent assignments (perhaps unrecognized) remain. A different sense of comparative incoherence is explored by Schervish, Seidenfeld, and Kadane (1997, 2002, 2003). Unlike Zynda (1996), which uses qualitative coherence orderings, they quantify degree of incoherence, utilizing the idea of vulnerability to Dutch books. Intuitively, having a degree of belief of 0.1 in $p \vee \sim p$ is "more incoherent" than believing it to degree 0.9. At the same stakes (say, $1), one would lose more by regarding 10¢ as a fair price for a bet on $p \vee \sim p$ than 90¢ (assuming, as is usual, that one is indifferent between buying and selling bets at those prices). The problem is to provide some shared measure of vulnerability to loss where one is betting on a number of propositions simultaneously, perhaps at different stakes. Schervish, Seidenfeld, and Kadane do this by defining the "worst" Dutch book that can be made against an incoherent agent, given a fixed "escrow" of funds (defined in various ways), and using this as a measure of degree of incoherence. Staffel (2015) critiques both these approaches and provides an alternative in terms of distance measures.

References

Allais, M. (1952/1979) The Foundations of a Positive Theory of Choice Involving Risk and a Criticism of the Postulate and Axioms of the American School. In Allais, M. and Hagen, O. (eds.) *Expected Utility Hypotheses and Allais' Paradox*. pp. 27–145. Dordrecht: Reidel.

Armendt, B. (1980) Is There a Dutch Book Argument for Probability Kinematics? *Philosophy of Science*. 47. pp. 583–8.

Brier, G.W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*. 78. pp. 1–3.

de Finetti, B. (1937/1964) Foresight: Its Logical Laws, Its Subjective Sources. In Kyburg, H. E., Jr. and Smokler, H. E. (eds.) *Studies in Subjective Probability*. Huntington, NY: Krieger.

Diaconis, P. and Zabell, S. (1982) Updating Subjective Probability. *Journal of the American Statistical Association*. 77. pp. 822–30.

Earman, J. (1992) *Bayes or Bust?* Cambridge, MA: MIT Press.

Eells, E. (1990) Bayesian Problems of Old Evidence. In Savage, C. W. *Scientific Theories*. Minnesota Studies in the Philosophy of Science. Vol. XIV. pp. 205–23. Minneapolis, MN: University of Minnesota Press.

Eriksson, L. and Hájek, A. (2007) What Are Degrees of Belief? *Studia Logica*. 86. July (Formal Epistemology I). pp. 185–215.

Fishburn, P. (1988) *Nonlinear Preference and Utility Theory*. Baltimore, MD: The Johns Hopkins University Press.

Garber, D. (1983) Old Evidence and Logical Omniscience. In Earman, J. (ed.) *Testing Scientific Theories*. Minnesota Studies in the Philosophy of Science. Vol. X. pp. 99–131. Minneapolis, MN: University of Minnesota Press.

Glymour, C. (1980) *Theory and Evidence*. Princeton, NJ: Princeton University Press.

Hájek, A. (2003) What Conditional Probability Could Not Be. *Synthese*. 137. pp. 273–323.

Hájek, A. (2005) Scotching Dutch Books? *Philosophical Perspectives*. 19 (Issue on Epistemology). pp. 139–51.

Hájek, A. (2008) Arguments for—or Against—Probabilism? *British Journal for the Philosophy of Science*. 59. 4. [Reprinted in Huber and Schmidt-Petri (2009). pp. 229–51.]

Hájek, A. (2009) Dutch Book Arguments. In Anand, P., Pattanaik, P., and Puppe, C. (eds.) *The Oxford Handbook of Rational and Social Choice*. pp. 173–95. Oxford: Oxford University Press.

Horowitz, T. (2006) Making Rational Choices When Preferences Cycle. In Horowitz, T. (ed. J. Camp) *The Epistemology of A Priori Knowledge*. Ch. 2. Oxford: Oxford University Press.

Howson, C. and Urbach, P. (1989) *Scientific Reasoning: The Bayesian Approach*. La Salle, IL: Open Court.

Huber, F. and Schmidt-Petri, C. (eds.) (2009) *Degrees of Belief*. London: Springer.

Jeffrey, R. (1965) *The Logic of Decision*. Chicago, IL: University of Chicago Press. 2nd edition, 1983.

Jeffrey, R. (1983/1992) Bayesianism with a Human Face. In Earman, J. (ed.) *Testing Scientific Theories*. Minnesota Studies in the Philosophy of Science. Vol. X. pp. 133–56. Minneapolis, MN: University of Minnesota Press. (Reprinted in Jeffrey, R. *Probability and the Art of Judgment*. pp. 77–107. Cambridge: Cambridge University Press.)

Jeffrey, R. (1988/1992) Conditioning, Kinematics and Exchangeability. In Skyrms, B. and Harper, W. (eds.) *Causation, Chance, and Credence*. Vol. 1. Boston, MA: Kluwer. pp. 221–55. (Reprinted in Jeffrey, R. *Probability and the Art of Judgment*. pp. 117–53. Cambridge: Cambridge University Press.)

Jeffrey, R. (1991/1992) Postscript 1991: New Explanation Revisited. In Jeffrey, R. *Probability and the Art of Judgment*. pp. 103–107. Cambridge: Cambridge University Press.

Jeffrey, R. (1992) *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.

Joyce, J. (1998) A Non-Pragmatic Vindication of Probabilism. *Philosophy of Science*. 65. pp. 575–603.

Joyce, J. (2009) Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Huber, F. and Schmidt-Petri, C. (eds.) *Degrees of Belief*. pp. 263–97.

Kahneman, D. and Tversky, A. (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*. 47. pp. 263–91.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kemeny, J. (1955) Fair Bets and Inductive Probabilities. *Journal of Symbolic Logic*. 20. pp. 263–73.

Kuhn, T. (1970) *The Structure of Scientific Revolutions*. 2nd edition. Chicago, IL: University of Chicago Press.

Lehman, R. (1955) On Confirmation and Rational Betting. *Journal of Symbolic Logic*. 20. pp. 251–62.

Leitgeb, H. and Pettigrew, R. (2010a) An Objective Justification of Bayesianism I: Measuring Inaccuracy. *Philosophy of Science*. 77. pp. 201–35.

Leitgeb, H. and Pettigrew, R. (2010b) An Objective Justification of Bayesianism II: The Consequences of Minimizing Inaccuracy. *Philosophy of Science*. 77. pp. 236–72.

Levi, I. (1987) The Demons of Decision. *The Monist*. 70. pp. 193–211.

Levi, I. (1991) Consequentialism and Sequential Choice. In Bacharach, M. and Hurley, S. (eds.) *Foundations of Decision Theory*. Oxford: Blackwell.

Lindley, D. (1982) Scoring Rules and the Inevitability of Probability. *International Statistical Review*. 50. pp. 1–26.

Maher, P. (1993) *Betting on Theories*. Cambridge: Cambridge University Press.

Maher, P. (2002) Joyce's Argument for Probabilism. *Philosophy of Science*. 96. pp. 73–81.

Ramsey, F. (1926/1990) Truth and Probability. In Ramsey, Frank (ed. D. H. Mellor) *Philosophical Papers*. pp. 52–109. Cambridge: Cambridge University Press.

Roseveare, N. (1982) *Mercury's Perihelion from Leverrier to Einstein*. Oxford: Oxford University Press.
Savage, L. (1954) *The Foundations of Statistics*. New York, NY: Wiley. [Revised and enlarged edition 1973. New York: Dover].
Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (1997) Two Measures of Incoherence. Tech. Report #660. Department of Statistics, Carnegie Mellon University, Pittsburgh, PA.
Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (2002) Measuring Incoherence. *Sankhyā: The Indian Journal of Statistics.* Series A. 64. 3. Part 1 pp. 561–87.
Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (2003) Measures of Incoherence. In Bernardo, J. M. et al. (eds.) *Bayesian Statistics. 7*. Oxford: Oxford University Press.
Schick, F. (1986) Dutch Books and Money Pumps. *Journal of Philosophy*. 83. pp. 112–19.
Shimony, A. (1955) Coherence and the Axioms of Confirmation. *Journal of Symbolic Logic*. 20. pp. 1–28.
Skyrms, B. (1984) *Pragmatics and Empiricism.* New Haven, CT: Yale University Press.
Skyrms, B. (1987) Dynamic Coherence and Probability Kinematics. *Philosophy of Science*. 54. pp. 1–20.
Skyrms, B. and Harper, W. (eds.) (1988) *Causation, Chance, and Credence*. Vol. 1. Boston, MA: Kluwer.
Staffel, J. (2015) Measuring the Overall Incoherence of Credence Functions. *Synthese*. 192. pp. 1467–93.
Teller, P. (1973) Conditionalization and Observation. *Synthese*. 26. pp. 218–58.
van Fraassen, B. (1983) Calibration: A Frequency Justification for Personal Probability. In Cohen, R. S. and Laudan, L. (eds.) *Physics, Philosophy, and Psychoanalysis*. pp. 295–319. Dordrecht: Reidel.
van Fraassen, B. (1986) A Demonstration of the Jeffrey Conditioning Rule. *Erkenntnis*. 24. pp. 17–24.
van Fraassen, B. (1988) The Problem of Old Evidence. In Austin, D. F. (ed.) *Philosophical Analysis*. pp. 153–65. Boston: Kluwer Academic Publishers.
van Fraassen, B. (1989) *Laws and Symmetry*. Oxford: Oxford University Press.
van Fraassen, B. (1990) Rationality Does Not Require Conditionalization. In Ullmann-Margalit, E. (ed.) *The Israel Colloquium: Studies in History, Philosophy, and Sociology of Science*. Vol. 5. Dordrecht: Kluwer.
von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.
Wagner, C. G. (1997) Old Evidence and New Explanation. *Philosophy of Science*. 64. pp. 677–91.
Wagner, C. G. (1999), Old Evidence and New Explanation II. *Philosophy of Science*. 66. pp. 283–8.
Wagner, C. G. (2001) Old Evidence and New Explanation III. *Philosophy of Science*. 68. 3. Supplement: Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers. September. pp. S165–S175.
Wagner, C. G. (2002) Probability Kinematics and Commutativity. *Philosophy of Science*. 69. pp. 266–78.
Williamson, J. (1999) Countable Additivity and Subjective Probability. *British Journal for the Philosophy of Science*. 50. pp. 401–16.
Zynda, L. (1995) Old Evidence and New Theories. *Philosophical Studies*. 77. pp. 67–95.
Zynda, L. (1996) Coherence as an Ideal of Rationality. *Synthese*. 109. pp. 175–216.
Zynda, L. (2000) Representation Theorems and Realism about Degrees of Belief. *Philosophy of Science*. 67. pp. 45–69.

CHAPTER 18

BAYESIANISM VS. FREQUENTISM IN STATISTICAL INFERENCE

JAN SPRENGER

Motivation: The Discovery of the Higgs Particle

Bayesianism and frequentism are the two grand schools of statistical inference, divided by fundamentally different philosophical assumptions and mathematical methods. In a nutshell, **Bayesian inference** is interested in the credibility of a hypothesis given a body of evidence, whereas frequentists focus on the reliability of the procedures that generate their conclusions. More exactly, a **frequentist inference** is valid if in the long run, the underlying procedure rarely leads to a wrong conclusion.

To better describe the scope and goals of these approaches, I follow Royall (1997, 6) in his distinction of three main questions in statistical analysis :

1. What should we *believe*?
2. What should we *do*?
3. When do data count as *evidence* for a hypothesis?

These questions are closely related, but distinct. Bayesians focus on the first question—rational belief—because for them, scientific hypotheses are an object of personal, subjective uncertainty. Therefore, Bayesian inference is concerned with the question of how data should change our degree of belief in a hypothesis. Consequently, Bayesians answer the second and third questions—what are rational decisions and good measures of evidence?—within a formal model of rational belief provided by the probability calculus.

Frequentists are united in rejecting the use of subjective uncertainty in the context of scientific inquiry. Still, they disagree considerably on the foundations of statistical inference. Behaviorists such as Jerzy Neyman and Egon Pearson build their statistical framework on reliable decision procedures, thus emphasizing the second question, while others, such as Ronald A. Fisher or Deborah Mayo, stress the relevance of (post-experimental) evidential assessments.

The purpose of this chapter is to make the reader understand the principles of the two major schools of statistical inference—Bayesianism and frequentism—and to recognize the scope, limitations, and weak spots of either approach. Notably, the divergences between the two frameworks can have a marked impact on the assessment of scientific findings. On 4 July, 2012, CERN (European Center for Nuclear Research) at Geneva surprised the public by the announcement of the discovery of the *Higgs boson*—a particle in the Standard Model of modern physics, which had been sought since 1964. Since the discovery of the Higgs boson proved the existence of a particular mechanism for breaking electroweak symmetry, this discovery was of extreme importance for particle physics.

In their statistical analysis, the researchers at CERN reasoned in a frequentist spirit: under the assumption that the Higgs boson does not exist, the experimental results deviate more than five standard deviations from the expected value. Since such an extreme result would occur by chance only once in two million times, the statisticians (and the press department) concluded that the Higgs boson had indeed been discovered.

This analysis sparked a vivid debate between Bayesian and frequentist statisticians. The well-known Bayesian statistician Tony O'Hagan sent an email to the newsletter of the International Society for Bayesian Analysis (ISBA) the entire statistical analysis in which was heavily attacked:

> We know from a Bayesian perspective that this [frequentist evidence standard, J.S.] only makes sense if (a) the existence of the Higgs boson...has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. Neither seems to be the case…. Are the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is? (O'Hagan 2012)

O'Hagan's message prompted a vivid exchange in the ISBA forum, with prominent statisticians and many particle physicists taking part in the discussion. In the first place, the debate concerned a specific standard of evidence, but since the notion of strong evidence depends on the chosen statistical framework, it quickly developed into a general dispute about the merits of Bayesian and frequentist statistics. Thus, the discovery of the Higgs particle exemplifies how the interpretation of a fundamental scientific result depends on methodological issues about statistical inference. Such cases are not limited to particle physics: they occur in every branch of science where statistical methods are used, and include issues as applied as the admission process for medical drugs.

Statistical methodology is thus a significant topic for philosophy, science, and public policy. In this chapter, we focus on how statistical evidence should be interpreted. This is not only the most contested ground between Bayesians and frequentists, but also utterly relevant for statisticians, experimenters, and scientific policy advisors. The chapter is structured as follows: Section 18.1 summarizes the principles of Bayesian inference. Sections 18.2 and 18.3 contrast behavioral and evidential interpretations of frequentist tests. Section 18.4 deals with the notorious p-values. Section 18.5 discusses confidence intervals as an alternative to significance tests and p-values whereas 18.6 deals with Mayo's error-statistical approach. Section 18.7 briefly exposes the optional stopping problem, and Section 18.8 concludes with a general discussion.

18.1 Bayesian Inference

Bayesian reasoners interpret probability as rational degree of belief. That is, an agent's system of degrees of belief is represented by a probability function $p(\cdot)$, and $p(H)$ quantifies his or her degree of belief that hypothesis H is true. These degrees of belief can be changed in the light of incoming information. The degree of belief in hypothesis H after learning evidence E is expressed by the conditional probability of H given E, $p(H|E)$:

Bayesian Conditionalization: The rational degree of belief in a proposition H after learning E is the conditional probability of H given E: $p_{\text{new}}(H) = p(H|E)$.[1]

$p(H)$ and $p(H|E)$ are called the **prior probability** and **posterior probability** of H. They can be related by means of **Bayes' Theorem**:

$$p_{\text{new}}(H) := p(H|E) = p(H)\frac{p(E|H)}{p(E)} = \left(1 + \frac{p(\neg H)}{p(H)} \cdot \frac{p(E|\neg H)}{p(E|H)}\right)^{-1} \tag{18.1}$$

The terms $p(E|H)$ and $p(E|\neg H)$ are called the **likelihoods** of H and $\neg H$ on E, that is, the probability of the observed evidence E under a specific hypothesis, in this case H or $\neg H$.

The label "Bayesian inference" usually refers to the conjunction of the following principles:

- The representation of subjective degrees of belief in terms of probabilities.
- The use of Bayesian Conditionalization for rationally revising one's degrees of belief.
- The use of the posterior probability distribution for assessing evidence, accepting hypotheses and making decisions.

However, not all Bayesians agree with these principles. Carnap's (1950) system of logical probability and Jeffreys' (1939) objective priors violate the first principle. Conditionalization is rejected by Bayesians who accept the Principle of Maximum Entropy (Williamson 2010). In this chapter, however, the focus is on the standard subjectivist position in Bayesian statistics that is built on the conjunction of these three principles.[2]

A consequence of Bayesianism that is of particular importance in statistical inference is the

> **Likelihood Principle (LP):** Consider a statistical model $\mathcal{M}$ with a set of probability measures $p(\cdot|\theta)$ parametrized by $\theta \in \Theta$. Assume we conduct an experiment $\mathcal{E}$ in $\mathcal{M}$. Then, all evidence about θ generated by $\mathcal{E}$ is contained in the *likelihood function* $p(x|\theta)$, where the observed data x are treated as a constant. (Birnbaum 1962; Berger and Wolpert 1984).[3]

To clarify, the **likelihood function** takes as argument the parameters of a statistical model, yielding the probability of the actually observed data under those parameter values. In

[1] See the chapter by Zynda in this volume (17) on the subjective interpretation of probability for a defense of Conditionalization, and for arguments that degrees of belief should satisfy the probability calculus.

[2] See Howson and Urbach (2006) for a philosophical introduction, and Bernardo and Smith (1994) for a more mathematically oriented treatment.

[3] We follow the convention of using capital letters for random variables and regular letters for their realizations.

particular, the LP entails that the probability of outcomes which have not been observed does not matter for the statistical interpretation of an experiment.

From the perspective of Bayes' Theorem, all that is needed to update a prior to a posterior is the likelihood of H and $\neg H$ given the observed data. In a statistical inference problem, this corresponds to the probability of the data x under various values of the unknown parameter θ. Therefore, it is absolutely logical that a subjective Bayesian endorses the LP.

As Birnbaum (1962) showed in a celebrated paper, the Likelihood Principle can be derived from two different and more basic principles: Sufficiency and Conditionality. We begin with the first one. A statistic (i.e., a function of the data X) $T(X)$ is *sufficient* if the distribution of the data X does not depend on the unknown parameter θ, conditional on T. In other words, sufficient statistics are compressions of the data set that do not lose any relevant information about θ. An example is an experiment about the bias of a coin. Assuming that the tosses are independent and identically distributed, the overall number of heads and tails is a sufficient statistics for an inference about the bias of the coin. Thus, we can neglect the precise order in which the results occurred. Formally, the **Sufficiency Principle** states that any two observations x_1 and x_2 are evidentially equivalent with regard to the parameter of interest θ as long as $T(x_1) = T(x_2)$ for a sufficient statistic T. Therefore, the principle is usually accepted by Bayesians and frequentists alike.

The **Conditionality Principle** is more controversial: it states that evidence gained in a probabilistic mixture of experiments is equal to the evidence in the actually performed experiment. In other words, if we throw a die to decide whether experiment $\mathcal{E}_1$ is conducted (in case the die comes up with an odd number) or experiment $\mathcal{E}_2$ (even number) and we throw a six, then the evidence from the overall experiment $\mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2$ is equal to the evidence from $\mathcal{E}_2$. Frequentists usually reject Conditionality since their measures of evidence take the entire sample space into account.[4]

According to many, it is the task of science to state the evidence for hypotheses of interest, rather than reporting degrees of belief in their truth. To address this challenge, the Bayesian needs a **measure of evidence**, that is, a numerical representation of the impact of the data on the hypotheses of interest. A particular measure is used almost universally: the **Bayes factor**, that is, the ratio of prior and posterior odds between hypothesis $H_0 : \theta \in \Theta_0$ and the alternative $H_1 : \theta \in \Theta_1$ conditional on data x (Kass and Raftery 1995).

$$B_{01}(x) := \frac{p(H_0|x)}{p(H_1|x)} \cdot \frac{p(H_1)}{p(H_0)} = \frac{\int_{\theta\in\Theta_0} p(x|\theta)p(\theta)d\theta}{\int_{\theta\in\Theta_1} p(x|\theta)p(\theta)d\theta}.^{5} \tag{18.2}$$

Thus, for two composite hypotheses H_0 and H_1, the Bayes factor can be written as the ratio of the integrated likelihoods, weighted with the prior plausibility of the individual hypotheses.

The Bayes factor is appealing for several reasons. Crucially, one can derive the posterior probability of a hypothesis H when one knows its prior $p(H)$ and the Bayes factor of H vs. $\neg H$. In the case of simple point hypotheses $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, the Bayes

[4] See Section 18.7. A thorough discussion of these principles goes beyond the scope of this chapter although some issues about Conditionality also are addressed also in the section on optional stopping. Recently, Mayo (2010) has challenged Birnbaum's proof of the LP from the Sufficiency and Conditionality principles.

factor reduces to the **likelihood ratio** $L(x, H_0, H_1) = p(x|H_0)/p(x|H_1)$, which possesses some interesting optimality properties as a measure of evidence (Lele 2004).

Some statisticians and methodologists use $L(x, H_0, H_1)$ as a contrastive measure of evidence without using the subjective probability interpretation. The reason is that they have doubts about whether subjective degrees of belief should be used in quantifying statistical evidence. These **likelihoodists** answer the third grand question—measuring evidence—by means of the thesis that L, or appropriate amendments thereof, provide the best measure of evidence (e.g., Royall 1997; Lele 2004). The likelihoodist framework, sometimes also called "Bayesianism without priors", has been anticipated by Hacking (1965) and elaborated by Edwards (1972) and most recently by the methodologist and biostatistician Richard Royall (1997).

Since likelihoodists and Bayesians agree on a lot of foundational issues—e.g., both camps accept the LP and raise similar objections against frequentism—I do not give a separate treatment of this approach. Certainly, it is conceptually and foundationally appealing, especially because it seems to do justice to the idea that statistical evidence is objective. However, likelihood-based inference is hard to implement in practice if the inference problems involve composite hypotheses and nuisance parameters. In those cases, computing $L(x, H_0, H_1)$ seems to require calculation of the marginal likelihoods, and thus, prior weights over the elements of the hypothesis space: $p(x|H_0) = \int_{\theta \in \Theta_0} p(x|\theta)p(\theta)$. So the likelihoodist either has to compromise the objectivity of her approach, effectively becoming a Bayesian, or to take refuge in other measures of evidence, such as conditional likelihood ratios. Royall (1997, ch. 7) discusses several *ad hoc* techniques that take care of important applications, but the fundamental philosophical problem remains.

18.2 Frequentism: Neyman and Pearson's Behavioral Approach

In the 19th century, probability theory gradually extended its scope from games of chance to questions of data analysis in science, industry and public administration. This was perhaps the birth hour of modern statistics. Yet, statistics was not as strongly linked to inductive inference as it is today. For instance, eminent statistician and biologist Francis Galton conceived of statistics as a *descriptive* tool for meaningfully compressing data, and summarizing general trends (cf. Aldrich (6) in this volume (Aldrich 2016)). Moreover, in those days the *nomothetic ideal*—to strive for certainty, for invariable, deterministic laws—had a great impact on scientific practice. Probability was used to quantify the lack of precision in measurement and conclusions, but not as a meaningful part of scientific theorizing (see the contributions in Krüger et al. 1987).

This attitude changed at the beginning of the 20th century with the groundbreaking discoveries of statisticians such as Karl Pearson and William Gosset ("Student"). Pearson discovered the χ^2-test (1900) for testing the goodness of fit between a hypothesized distribution and the observed data, Gosset discovered the t-test (1908) for making inferences about the mean of a Normally distributed population. These techniques, which are still widely used today, were invented in response to applied research questions and

mark the transition from **descriptive** to **inferential statistics**. Statistics became a discipline concerned with making inferences about a parameter of interest, predictions and decisions, rather than just summarizing data.

Given the aforementioned nomothetic, objectivist ideals, many scientists had issues with the Bayesian approach to probabilistic inference. After all, subjective degrees of belief are hard to measure, and apparently lack the impartiality and objectivity of scientific findings. The great statistician Ronald A. Fisher (1935, pp. 6–7) even spoke of "mere psychological tendencies, theorems respecting which are useless for scientific purposes". Fisher clarified that he did not believe Bayesian reasoning to be logically invalid, but that there is rarely any reliable information on which a non-arbitrary prior probability distribution could be based.

The need to develop a coherent non-Bayesian theory of probabilistic inference was therefore sorely felt. One famous answer was given by the British statisticians Jerzy Neyman and Egon Pearson who connected statistical analysis to rational decision-making. In their groundbreaking 1933 paper, they developed a genuinely frequentist theory of hypothesis testing: statistical tests should be constructed as to minimize the relative frequency of wrong decisions in a hypothetical series of repetitions of the test. In particular, Neyman and Pearson linked the interpretation of statistical experiments tightly to their design.

An example may illustrate their approach. Suppose that we must decide whether a medical drug is going to make it to the next phase in a cumbersome and expensive drug admission process. Of course, we would not like to admit a drug that is no better than existing treatments in terms of efficacy, side effects, costs, and so on. On the other hand, we do not want to erroneously eliminate a superior drug from the admission process. These are the two possible kinds of errors, commonly called **type I** and **type II error**.

For making a sound decision, Neyman and Pearson suggest the following route: first, the scientist chooses a default or **null hypothesis** H_0 for which a type I error rate is fixed. In medicine, the null usually states that the new treatment brings no improvement over the old one. After all, admitting an inefficient or even harmful drug is worse than foregoing a more effective treatment—at least from a regulatory point of view. By contrast, the **alternative** H_1 states that the drug is a genuine improvement. While a type I error corresponds to erroneous rejection of the null hypothesis, a type II error stands for erroneous acceptance of the null.

Conventionally, acceptable type I error rates are set at a **level** of 5%, 1% or 0.1%, although Neyman and Pearson insist that these levels have no special meaning, and that striking the balance between type I and type II error rates is a highly context-sensitive endeavor. In good frequentist spirit, Neyman and Pearson devise a decision procedure such that (i) in not more than 5%/1%/0.1% of all cases where the null hypothesis is true, it will be rejected; (ii) the **power** of the test—its ability to discern the alternative when it is true—is maximal for the chosen level of the test. In other words, given a fixed type I error rate (e.g., 1%), we design the test such that the type II error rate is minimized. Then, we are rational in following the test procedure because of its favorable long-run properties:

> [...] we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false. (Neyman and Pearson 1967, p. 142)

But how do we find the optimal test? For the case of two point hypotheses ($\theta = \theta_0$ vs. $\theta = \theta_1$) being tested against each other, Neyman and Pearson have proved an elegant result:

Fundamental Lemma of Neyman and Pearson (1933): When testing two point hypotheses against each other, the most powerful test at any level α is the *likelihood ratio test*. This is a test T for which there is a $C(\alpha) \in \mathbb{R}$ such that for data x:

$$T(x) = \begin{cases} \text{accept } H_0 & \text{if } L = \frac{p(x|\theta=\theta_0)}{p(x|\theta=\theta_1)} \geq C(\alpha) \\ \text{reject } H_0 & \text{if } L = \frac{p(x|\theta=\theta_0)}{p(x|\theta=\theta_1)} < C(\alpha) \end{cases} \tag{18.3}$$

Hence, the uniformly optimal test in Neyman and Pearson's sense depends on the weight of evidence as measured by the likelihood ratio L. If L strongly favors H_1 over H_0, we will reject the null, otherwise we will accept it. This result is at the bottom of a powerful mathematical theory of hypothesis testing and has greatly influenced statistical practice. However, Neyman and Pearson's approach has been attacked from a methodological point of view: according to Fisher, such tests are clever decision tools, but miss the point of scientific research questions. The next section explains this criticism.

18.3 Frequentism: Significance Tests and Fisher's Disjunction

The second grand tradition in frequentist statistics emerged with Ronald A. Fisher, eminent geneticist and statistician, who violently opposed Neyman and Pearson's behavioral, decision-theoretic approach. In determining an acceptable type I error rate, Neyman and Pearson implicitly determine the severity of an error, thereby imposing a decision-theoretic utility structure on the experiment in question. Fisher argued, however, that

> in the field of pure research no assessment of the cost of wrong conclusions ... can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence. (Fisher 1935, pp. 25–6)

Two arguments are implied here. First, we cannot quantify the utility that correctly accepting or rejecting a hypothesis will eventually have for the advancement of science. The far-reaching consequences of such a decision lie beyond our horizon. Second, statistical hypothesis tests should state the *evidence* for or against the tested hypothesis: a scientist is interested in whether she has reason to believe that a hypothesis is *true* or *false*. Her judgment should not be obscured by the practical consequences of working with this rather than that hypothesis. Therefore, Neyman-Pearson tests may be helpful in industrial quality control and other applied contexts, but not in finding out the truth about a scientific hypothesis.[6]

For Fisher (1956) himself, the purpose of statistical analysis consisted in assessing the relation of a (null) hypothesis to a body of observed data. That hypothesis usually stands for there being no effect of interest, no causal relationship between two variables, etc. In other words, the null denotes the absence of a phenomenon to be demonstrated.

[6] The classification of Neyman-Pearson tests as purely behavioral is not without contention. Following the representation theorems in Savage (1962), one might link Neyman-Pearson tests to a general theory of belief attitudes and rational decision-making. Romeijn (2010, section 9) also investigates an embedding of Neyman-Pearson tests into Bayesian statistics.

Then, the null is tested for being compatible with the data—notably, without considering explicit alternatives. This is called a **significance test**. Thus, Fisher's approach is essentially asymmetric: while a "rejection" strongly discredits the null hypothesis, an "acceptance" just means that the facts have failed to disprove the null. By contrast, Neyman-Pearson tests with sufficiently high power are essentially symmetric in the interpretation of the outcome.

The basic rationale of significance testing, called "Fisher's Disjunction" by Hacking (1965), is as follows: a very unlikely result undermines the (objective) tenability of the null hypothesis.

> "either an exceptionally rare chance has occurred, or the theory [=the null hypothesis] is not true." (Fisher 1956, p. 39)

The occurrence of such an exceptionally rare chance has both epistemological and practical consequences: first, the null hypothesis is rendered "objectively incredible" (Spielman 1974, p. 214), second, the null should be treated as if it were false. Naturally, this judgment is not written in stone, but may be overturned by future evidence. Here is a graphical representation of Fisher's main idea.

> p(Data|Null Hypothesis) is low.
>
> Data is observed.
>
> Null Hypothesis is discredited.

Notably, Fisher's ideas are close to Popper's falsificationism, albeit with more inductivist inclinations. They both agree the only purpose of an experiment is to "give the facts a chance of disproving the null hypothesis" (Fisher 1925, p. 16). They also agree that failure to reject a hypothesis does not conclude positive evidence for the tested (null) hypothesis. But unlike Popper (1934/1959), Fisher aims at experimental and statistical *demonstrations* of a phenomenon.

The above scheme of inference (cf. also Gillies 1971) has been criticized frequently. Hacking (1965, pp. 81–2) has pointed out that the explication of the term "exceptionally rare chance" inevitably leads to trouble. A prima facie reading of Fisher's objection cited above quote seems to suggest that the chance of the observed event must be exceptionally low compared to other events that could have been observed. But in that case, some statistical hypotheses could never be tested. For instance, a uniform distribution over a finite set of events assigns equal likelihood to all observations, so there is no exceptionally rare chance. How should we test—and possibly reject—such a hypothesis?

To expand on this point, imagine that we are now testing the hypothesis that a particular coin is fair. Compare now two series of independent and identically distributed tosses: "HTTHTTTHHH" and "HHHHHHHHHH". The probability of both events under the null is the same, namely $(1/2)^{10} = 1/1024$. Still, the second series, but not the first, seems to strongly speak against the null. Why is this the case? Implicitly, we have specified *the way in which the data are exceptional*: namely, we are interested in the propensity θ of the coin to come up tails. Since T, the number of tails, is a sufficient statistic with respect to θ, we can restrict our attention to the value of T. Then, $\{T = 0\}$ is indeed a much less likely event than $\{T = 5\}$ (cf. Royall 1997, ch. 3).

It seems that we cannot apply significance tests without an implicit specification of alternative hypotheses; here: that the coin is biased toward tails. Spielman (1974) further

presses this point in an extended logical analysis of significance testing: inferring from an unlikely result to the presence of a significant effect *presupposes* that the observed result is much more likely under an implicitly conceived alternative than under the null. Otherwise we would have no reason to appraise that effect. Indeed, modern frequentist approaches, such as Mayo's (1996) error-statistical account, take this into account by explicitly setting up statistical inference in a contrastive way. That is, testing always occurs with respect to a direction of departure from the tested hypothesis.

However, does this modification suffice to save the logic of significance testing? Consider a blood donor who is routinely tested for an HIV infection. Let the null hypothesis state that the donor has not contracted HIV. The test returns the correct result in 99% of all cases, regardless of whether an HIV infection is present or not. Now, the test returns a positive result. Under the null, this certainly constitutes an exceptionally rare chance whereas under the alternative, it is very likely. Should the donor now be convinced that he has contracted HIV, given a general HIV prevalence of 0.01% in the population?

A Bayesian calculation yields, perhaps surprisingly, that he should still be quite certain of *not* having contracted HIV:

$$
\begin{aligned}
&p(\text{HIV contraction}|\text{positive test}) \\
&\quad = \left(1 + \frac{p(\text{positive test}|\text{no contraction})}{p(\text{positive test}|\text{HIV contraction})} \frac{p(\text{no contraction})}{p(\text{HIV contraction})}\right)^{-1} \\
&\quad = \left(1 + \frac{0.01}{0.99} \frac{0.9999}{0.0001}\right)^{-1} \approx 0.01
\end{aligned}
$$

In other words, the evidence for a contraction is more than cancelled by the very low base rate of HIV infections in the relevant population. Therefore, straightfowardly rejecting the null hypothesis on the basis of a "significant" finding is no valid inference, even if the findings are likely under the alternative. Since the fallacy is caused by neglecting the base rates in the populations, it is called the **Base Rate Fallacy** (cf. Goodman 1999).

Thus, if a significance test is supposed to deliver a valid result, the null must not be too credible beforehand (cf. Spielman 1974, p. 225). But if we make all these restrictions to Fisher's proposal, it may be asked why we should not switch to a straight Bayesian approach. After all, both approaches involve judgments of prior credibility, and the Bayesian framework is much more systematic and explicit in making and revising such judgments, and in integrating various sources of information.[7]

The above criticisms show that significance testing is logically invalid. To rescue it, we have to make additional premises, some of which incorporate a Bayesian viewpoint. But if all this is right, why is significance testing such a widespread tool in scientific research? This question will be addressed in the next section.

[7] Spanos (2010, pp. 576–80) objects that properly conceptualized frequentist tests do not fall prey to the Base Rate Fallacy: frequentist hypotheses describe an unknown data-generating mechanism. Whereas the hypothesis of interest in the above example (whether the donor has contracted HIV) is just an event in a more general statistical model that describes contraction status and test results of the entire population.

18.4 Frequentism: p-values

Significance testing in the Fisherian tradition is arguably the most popular methodology in statistical practice. But there are important distinctions between Fisher's original view, discussed above, and the practice of significance testing in the sciences, which is a hybrid between the Fisher and the Neyman-Pearson school of hypothesis testing, and where the concept of the **p-value** plays a central role.

To explain these differences, we distinguish between a **one-sided** and a **two-sided testing problem**. The one-sided problem concerns the question of whether an unknown parameter is greater or smaller than a particular value ($\theta \leq \theta_0$ vs. $\theta > \theta_0$), whereas the two-sided testing problem (or point null hypothesis test) concerns the question of whether or not parameter θ is exactly equal to θ_0: $H_0 : \theta = \theta_0$. vs. $H_1 : \theta \neq \theta_0$. The two-sided test can be used for asking different questions: first, whether there is "some effect" in the data (if the null denotes the absence of a causal relationship), second, whether H_0 is a suitable proxy for $H_0 \vee H_1$, that is, whether the null is a predictively accurate idealization of a more general statistical model.

This use of hypothesis tests differs from Fisher's since he considered inferences within a parametric model primarily as a problem of **parameter estimation**, not of hypothesis testing (cf. Spielman 1978, p. 122). His method of significance testing was devised for testing hypotheses without considering alternative hypotheses. But owing to the problems mentioned in the previous section and the influence of Neyman and Pearson, modern significance tests require the specification of an alternative hypothesis. However, their interpretation is not behavioral, as Neyman and Pearson would require, but evidential, as Fisher would have requested.

The central concept of modern significance tests—the p-value—is now illustrated in a two-sided testing problem. Again, we want to infer the presence of a significant effect in the parameter θ if the discrepancy between data $x := (x_1, \ldots, x_N)$, corresponding to N realizations of an experiment, and null hypothesis $H_0 : \theta = \theta_0$ is large enough. Assume now that the variance σ^2 of the population is known. Then, one measures the discrepancy in the data x with respect to the postulated mean value θ_0 by means of the standardized statistic

$$z(x) := \frac{\frac{1}{N}\sum_{i=1}^{N} x_i - \theta_0}{\sqrt{N \cdot \sigma^2}} \tag{18.4}$$

We may re-interpret equation (18.4) as

$$z = \frac{\text{observed effect} - \text{hypothesized effect}}{\text{standard error}} \tag{18.5}$$

Determining whether a result is significant or not depends on the p-value or **observed significance level**, that is, the "tail area" of the null under the observed data. This value depends on z and can be computed as

$$p := p(|z(X)| \geq |z(x)|), \tag{18.6}$$

that is, as the probability of observing a more extreme discrepancy under the null than the one which is actually observed. Figure 18.1 displays an observed significance level $p = 0.072$

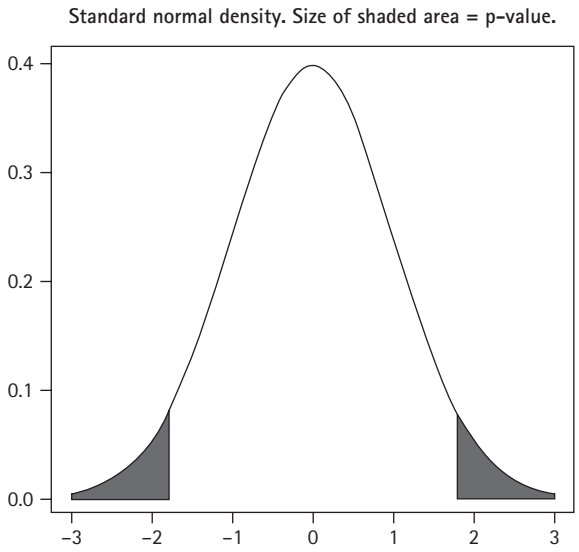


FIGURE 18.1 The probability density function of the null $H_0 : X \sim N(0, 1)$, which is tested against the alternative $H_1 : X \sim N(\theta, 1)$, $\theta \neq 0$. The shaded area illustrates the calculation of the p-value for observed data $x = 1.8$ ($p = 0.072$).

as the integral under the probability distribution function. For the frequentist practitioner, p-values are practical, replicable and objective measures of evidence against the null: they can be computed automatically once the statistical model is specified, and only depend on the sampling distribution of the data under H_0. Fisher interpreted them as "a measure of the rational grounds for the *disbelief* [in the null hypothesis] it augments" (Fisher 1956, 43).[8]

The virtues and vices of significance testing and p-values have been discussed at length in the literature, and it would go beyond the scope of this chapter to deliver a comprehensive discussion (see e.g., Cohen 1994; Harlow et al. 1997). The most important criticisms are discussed below and in Section 18.7 where the **sample space dependence** of frequentist inference will be thematized (cf. Hartmann and Sprenger 2011).

18.4.1 p-values and Posterior Probabilities

The arguably biggest problem with p-values is practical: many researchers are unable to interpret them correctly. Quite often, a low p-value (e.g., $p < 0.001$) is taken as the statement that the null hypothesis has a posterior probability smaller than that number (e.g., Oakes 1986; Fidler 2005). Of course, this is just an instance of the Base Rate Fallacy: subjects conflate the conditional probability of the evidence given the hypothesis, $p(E|H)$, with the

[8] See Romeijn (2010) for further exposition of an epistemic reading of frequentist statistics, including Fisher's *fiducial argument*.

conditional probability of the hypothesis given the evidence, $p(H|E)$. In other words, they conflate statistical evidence with rational degree of belief.

Despite persistent efforts to erase the Base Rate Fallacy, it continues to haunt statistical practitioners. Some have argued that this is an effect of the unintuitive features of the entire frequentist framework. For example, the German psychologist Gerd Gigerenzer (1993) argues that scientists are primarily interested in the tenability or credibility of a hypothesis, not in the probability of the data under the null. The question is then: how should we relate p-values to posterior probabilities? After all, a Bayesian and a frequentist analysis should agree when prior probability distributions can be objectively grounded.

It turns out that in the one-sided testing problem, p-values can often be related to posterior probability (Casella and Berger 1987, more on this in Section 18.6) whereas in the two-sided or point null testing problem, the two measures of evidence diverge. When the prior is uninformative, a low p-value may still entail a high posterior probability of the null. More precisely, Berger and Sellke (1987) show that the p-value is often proportional to a *lower bound* on the posterior probability of the null, thus systematically overstating the evidence against the null. This suggests a principal incompatibility between frequentist and Bayesian reasoning in the two-sided testing problem. We expand on this point in a later subsection when discussing Lindley's paradox.

18.4.2 p-values Vs. Effect Size

Another forceful criticism of p-values and significance tests concerns their relation to effect size. The economists Deirdre McCloskey and Stephen Ziliak have launched strong attacks against significance tests in a series of papers and books (McCloskey and Ziliak 1996; Ziliak and McCloskey 2008). Let us give their favorite example. Assume that we have to choose between two diet cures, based on pill A and pill B. Pill A makes us lose 10 pounds on average, with an average variation of 5 pounds.[9] Pill B makes us lose 3 pounds on average, with an average variation of 1 pound. Which one leads to more significant loss? Naturally, we opt for pill A because the effect of the cure is so much larger.

However, if we translate the example back into significance testing, the order is reversed. Assume the standard deviations are known for either pill. Compared to the null hypothesis of no effect at all, observing a three pounds weight loss after taking pill B is a more significant result evidence for the efficacy of that cure than observing a ten pounds weight loss after taking pill A:

$$z_A(10) = \frac{10-0}{5} = 2 \qquad\qquad z_B(3) = \frac{3-0}{1} = 3$$

Thus, there is a notable discrepancy between our intuitive judgment and the one given by the p-values. This occurs because statistical significance is supposed to be "a measure of the strength of the signal relative to background noise" (Hoover and Siegler 2008, p. 58). On this score, pill B indeed performs better than pill A, because of the favorable signal/noise ratio. But pace Ziliak and McCloskey, economists, businesspersons and policy-makers are

[9] The concept of "average variation" is intuitively explicated as the statistical concept of standard deviation, which is, for a random variable X, defined as $\sqrt{E[(X-E(X))^2]}$.

interested in the effect size, not the signal/noise ratio: they do not want to ascertain the presence of *some* effect, but to demonstrate a *substantial* effect, as measured by **effect size**.

This fundamental difference is, however, frequently neglected. By scrutinizing the statistical practice in the top journal *American Economic Review*, as well as by surveying the opinions of economists on the meaning of statistical significance, McCloskey and Ziliak derive the conclusion that most economists are unaware of the proper meaning of statistical concepts. In practice, "asterisking" prevails: e.g., in correlation tables, the most significant results are marked with an asterisk, and these results are the ones that are supposed to be real, big, and of practical importance. But an effect need not be statistically significant to be big and remarkable (like pill *A*), and a statistically significant effect can be quite small and uninteresting (like pill *B*).

18.4.3 p-values and Lindley's Paradox

The tension between effect size and statistical significance is also manifest in one of the most famous paradoxes of statistical inference, Lindley's Paradox. Classically, it is stated as follows:

> **Lindley's Paradox:** Take a Normal model $N(\theta, \sigma^2)$ with known variance σ^2 and a two-sided testing problem $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. Assume $p(H_0) > 0$, and any regular proper prior distribution on $\{\theta \neq \theta_0\}$. Then, for any testing level $\alpha \in [0, 1]$, we can find a sample size $N(\alpha)$ and independent, identically distributed data $x = (x_1, \ldots, x_N)$ such that
>
> 1. The sample mean $\bar{x}$ is significantly different from θ_0 at level α;
> 2. $p(H_0|x)$, that is, the posterior probability that $\theta = \theta_0$, is at least as big as $1 - \alpha$. (cf. Lindley 1957, p. 187)

In other words, a Bayesian and a frequentist analysis of a two-sided test may reach completely opposite conclusions. The reason is that the combination of statistical significance and large sample size (= high power) is highly misleading. In fact, as sample size increases, an ever smaller discrepancy from the null suffices to achieve a statistically significant result against the point null. The reader will thus be lured into believing that a "significant" result has substantial scientific implications although the effect size is very small. The high power of a significance test with many observations provides no protection against inferring to insignificant effects, quite to the contrary. Therefore, Lindley's Paradox lends forceful support to Ziliak and McCloskey's claim that statistical significance is a particularly unreliable guide to scientific inference.

This is not to say that all is well for the Bayesian: Assigning a strictly positive degree of belief $p(H_0) > 0$ to the point null hypothesis $\theta = \theta_0$ leads to a misleading and inaccurate representation of our subjective uncertainty. After all, $\theta = \theta_0$ is not much more credible than any value $\theta_0 \pm \varepsilon$ in its neighborhood. Therefore, assigning a strictly positive prior to H_0, instead of a continuous prior, seems unmotivated (cf. Bernardo 2012).

But if we set $p(H_0) = 0$, then for most priors (e.g., an improper uniform prior) the posterior probability distribution will peak not at the null value, but somewhere else. Thus, the apparently innocuous assumption $p(H_0) > 0$ has a marked impact on the result of the

Bayesian analysis. Attempts to consider it as a mathematical approximation of testing the hypothesis $H_0 : |\theta - \theta_0| < \varepsilon$ break down as sample size increases (cf. Berger and Delampady 1987).

The choice of prior probabilities, for H_0 as well as over the elements of H_1, is therefore a very sensitive issue in Lindley's Paradox. Quite recently, the Spanish statistician José M. Bernardo (1999, 2012) has suggested to replacing the classical Bayesian focus on posterior probability as a decision criterion with the Bayesian Reference Criterion (BRC), which focuses on the predictive value of the null in future experiments. This move avoids assigning strictly positive mass to a set of measure zero $\{\theta = \theta_0\}$ and reconciles Bayesian and frequentist intuitions to some extent. Sprenger (2013a) provides a more detailed discussion of this approach.

18.4.4 p-values and the Assessment of Research Findings

A methodological problem with p-values, stemming from their roots in Fisherian significance testing, is that insignificant results (= p-values greater than .05) have barely a chance of getting published. This is worrisome for at least two reasons: first, even a statistically insignificant result may conceal a big and scientifically relevant effect, as indicated by Ziliak and McCloskey; secondly, it prevents an appraisal of the evidence *in favor of the null hypothesis*. As a consequence, valuable resources are wasted because different research teams replicate insignificant results over and over again, not knowing of the efforts of the other teams. In addition, the frequentist provides no logic of inference for when an insignificant result supports the null, rather than just failing to reject it.

This asymmetry in frequentist inference is at the bottom of Ioannides' (2005) famous thesis that "most published research findings are false". Ioannides reasons that there are many false hypotheses that may be erroneously supported and yield a publishable research finding. If we test for significant causal relationships in a large set of variables, then the probability of a false positive report is, for type I and type II error rates α and β, normally larger than the probability that a true hypothesis be found. In particular, if R denotes the ratio of true to false relationships that are tested in a field of scientific inquiry and a "significant" causal relationship is found, then

$$p(\text{the supposed causal relationship is true}) = \frac{(1-\beta)\cdot R}{(1-\beta)\cdot R + \alpha} \tag{18.7}$$

This quantity is smaller than $1/2$ if and only if $R < \alpha/(1-\beta)$, which will typically be satisfied, given that $\alpha = .05$ is the standard threshold for publishable findings, and that most causal relationships that scientists investigate are not substantial. Thus, most published research findings are indeed artifacts of the data and plainly false—an effect that is augmented by the experimenter's bias in selecting and processing his or her data set.

This finding is not only a feature of scientific inquiry in general, but is specifically due to the frequentist logic of inference: the one-time achievement of a significant result is just not a very good indicator for the objective credibility of a hypothesis. Indeed, researchers often fail to replicate findings by another scientific team, and periods of excitement and subsequent disappointment are not uncommon in frontier science. The problems with frequentist inference affect the success of entire research programs.

18.5 Confidence Intervals as a Solution?

The above criticisms dealt severe blows to classical significance tests and the use of p-values. In the last decades, frequentists have therefore adapted their tools. Nowadays, they often replace significance tests with confidence intervals, allegedly a more reliable method of inference (e.g., Cumming and Finch 2005; Fidler 2005). Confidence intervals are interval estimators that work as follows. Let $C(\cdot,\cdot)$ be a subset of $\Theta \times \mathcal{X}$ for parameter space Θ and sample space $\mathcal{X}$. Then consider the set $C(\theta_0,\cdot)$ that comprises those (hypothetical) data points for which the hypothesis $\theta = \theta_0$ would *not* be rejected at the level α. In other words, $C(\theta_0,\cdot)$ contains the data points that are *consistent* with θ_0.

If we construct these sets for all possible values of θ, then we obtain a two-dimensional set C with $(\theta, x) \in \Theta \times \mathcal{X}$. Assume further that we observe data x_0. Now define the projection of C on the data x_0 by means of $C_{x_0} := \{\theta | x_0 \in C(\theta, x_0)\}$. This set $C_{x_0} \subset \Theta$ is called the **confidence interval** for parameter θ at level α, on the basis of data x_0.

Confidence intervals should not be understood in the literal sense that upon observing x_0, parameter θ lies in the interval C_{x_0} with probability $1-\alpha$. After all, the frequentist does not assign any posterior probability to the parameters of interest. Rather, the level of the confidence interval says something about the procedure used to construct it: in the long run, the observed data x will be consistent with the constructed intervals for θ in $100 \cdot (1-\alpha)\%$ of all cases, independent of the actual value of θ.

The advantage of confidence intervals over significance tests can be illustrated easily in the case of Lindley's Paradox. If we constructed a 95% confidence interval for θ, it would be a very narrow interval in the neighborhood of θ_0. Under the conditions of large sample size with low effect size, a confidence interval would avoid the false impression that the null was substantially mistaken.

However, confidence intervals do not involve a decision-theoretic component; they are interval estimators. If we take seriously that scientists want to conduct real *tests*, instead of estimating parameters, then confidence intervals cannot alleviate the worries with frequentist inference. Rather than solving the problem, they shift it, although they are certainly an improvement over naïve significance testing.

That said, confidence intervals rather fulfill the function of a consistency check than of inspiring trust in a specific estimate. They list the set of parameter values for which the data fall into the acceptance region at a certain level. This is in essence a pre-experimental perspective. But this does not warrant, post-experimentally, that the parameter of interest is "probably" in the confidence interval. Therefore, some frequentists are not happy with confidence intervals either. In recent years, the philosopher of statistics Deborah Mayo (1996) has tried to establish *degrees of severity* as superior frequentist measures of evidence. The next section is devoted to discussing her approach.

18.6 Mayo's Error-Statistical Account

In her 1996 book "Error and the Growth of Experimental Knowledge", Deborah Mayo works out a novel account of frequentist inference. Mayo's key concept, degrees of severity, combines Neyman and Pearson's innovation regarding the use of definite alternatives and

the concept of power with Fisher's emphasis on post-experimental appraisals of statistical evidence.

Mayo's model of inference stands in a broadly Popperian tradition: for her, it is essential to scientific method that a hypothesis that we appraise has been well probed (=severely tested). Why should passing a test count in favor of a hypothesis? When are we justified to rely on such a hypothesis?

Popper (1934/1959, p. 282) gave a skeptical reply to this challenge: he claimed that corroboration—the survival of past tests—is just a report of past performance and does not warrant any inference with regard to future expectations. Mayo wants to be more constructive and to entitle an *inference* to a hypothesis:

> evidence E should be taken as good grounds for H to the extent that H has passed a severe test with E. (Mayo 1996, p. 177)

Regarding the notion of what it means (for a statistical hypothesis) to pass a severe test, she adds:

> a passing result is a severe test of hypothesis H just to the extent that it is very improbable for such a result to occur, were H false (Mayo 1996, p. 178)

Notably, a null hypothesis which passes a significance test would, on Mayo's account, not necessarily count as being severely tested. For example, in tests with low sample size, the power of the test would typically be small, and even a false null hypothesis would probably pass the test. This is one of the reasons why she insists that hypotheses are *always tested against definite alternatives*. By means of quantifying how well a statistical hypothesis has been probed, we are entitled to inferences about the data-generating process. This exemplifies the basic frequentist idea that statistical inferences are valid if they are generated by reliable procedures.

For Mayo, a hypothesis H is *severely tested with data* x if (S-1) the data agree with the hypothesis, and (S-2) with very high probability, the data would have agreed less well with H if H were false (Mayo and Spanos 2006, p. 329).[10]

We illustrate her approach with an example of a Normally distributed population $N(\theta, \sigma^2)$ with known variance σ^2. Assume that we want to quantify the **severity** with which the hypothesis $H_0 : \theta \leq \theta_0$ passes a test T with observed data x (against the alternative $H_1 : \theta > \theta_0$). First, we measure the discrepancy of the data from the hypotheses by means of the well-known statistics

$$z_{\theta_0}(X) = \sqrt{N}(\bar{X} - \theta_0)/\sigma.$$

z measures the distance of the data from H_0 *in the direction of* H_1 (cf. Mayo and Spanos 2006, p. 331): a large value of $\bar{X} - \theta_0$ yields large values of z and thus, evidence against the null. Then, the severity with which H_0 passes a test with data x is defined as the probability that $z_{\theta_0}(X)$ would have taken a higher value if the alternative $H_1 : \theta > \theta_0$ had been true.

[10] The precise meaning of (S-1) remains a bit unclear; Mayo and Spanos (2006, p. 336) say in passing that statistically insignificant results "agree" with the null. This definition may be contested, however: depending on the choice of the alternative, insignificant results may strongly discredit the null.

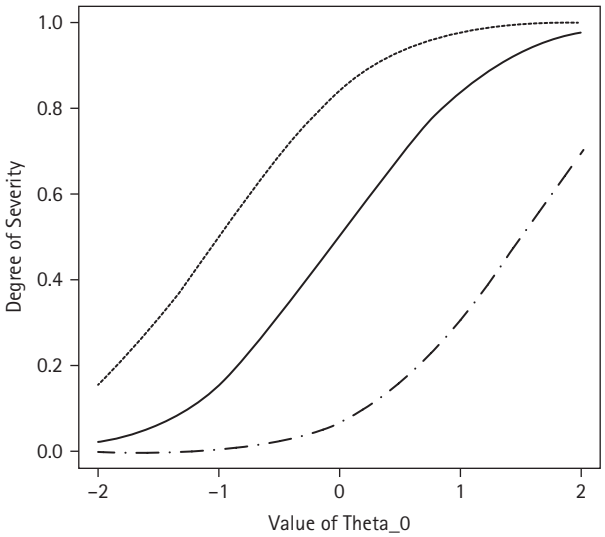


FIGURE 18.2 Inference about the mean θ of a normal population with variance $\sigma^2 = 1$. The three curves show the degrees of severity at which the hypothesis $\theta \leq \theta_0$ is accepted for three different data points. Dotted line: severity function for data $x = -1$, full line: $x = 0$, dashed line: $x = 1.5$.

Mathematically:[11]

$$\text{SEV}(\theta \leq \theta_0)(x, H_1) = p(z_{\theta_0}(X) > z_{\theta_0}(x); \theta > \theta_0). \tag{18.8}$$

As the alternative $H_1 : \theta > \theta_0$ comprises a large set of hypotheses that impose different sampling distributions on the z-statistic, there is an ambiguity in (18.8). Which element of H_1 should be used for calculating the probability of the right-hand side? To resolve this problem, Mayo observes that a lower bound on the test's severity is provided by calculating severity with respect to the hypothesis $\theta = \theta_0$. Thus, equation (18.8) becomes

$$\text{SEV}(\theta \leq \theta_0)(x, H_1) = p(z_{\theta_0}(X) > z_{\theta_0}(x); \theta = \theta_0). \tag{18.9}$$

This is, however, only half of the story. Mayo would also like to calculate at which severity the claim $\theta \leq \theta_0$ passes a test as a function of θ_0 when x is kept constant. Therefore, she calculates the **severity function** for θ_0 indicating which discrepancies from the null are warranted by the actual data, and which are not. Figure 18.2 gives an illustration.

The main merit of Mayo's approach consists in the systematiziation of the various intuitions, heuristics and rationales in frequentist statistics. Practice is often a hodgepodge of methods, inspired by ideas from both the Neyman-Pearson and the Fisherian school. In

[11] I modify the notation in Mayo and Spanos (2006) to some extent. However, I follow Mayo and Spanos in using the semicolon for separating event and hypothesis in the calculation of the degree of severity because for them, the difference from the vertical dash (and from conditional probability) carries philosophical weight.

particular, practitioners often combine decision procedures—calculating the power of a test, accepting/rejecting a null, etc.—with post-data evidential assessments, such as "hypothesis H_0 was rejected in the experiment (p=.026, power=.87)". Strictly speaking, this mix of Fisherian and Neyman-Pearson terminology is incoherent. With the error-statistical philosophy of inference and the concept of degree of severity, there is now a philosophical rationale underlying this practice: Mayo supplements Neyman and Pearson's pre-experimental methodology for designing powerful tests with a post-experimental measure of evidence. Therefore Mayo's approach is reasonably close to a lot of scientific practice carried out in the framework of frequentist statistics.

That said, there are also a number of problems for Mayo's approach, owing mainly to foundational problems that are deeply entrenched in the entire frequentist framework.

First, error statistics reduces the testing of composite hypotheses against each other (e.g., $\theta \leq \theta_0$ vs. $\theta > \theta_0$) to testing a hypothesis against that particular hypothesis which provides the most severe test (in this case, $\theta = \theta_0$). Thus, it may be asked whether degrees of severity really improve on a traditional frequentist or a likelihoodist analysis.

Secondly, there is a close mathematical relationship between degrees of severity and (one-sided) p-values. Both are derived from the cumulative distribution function of the z-statistic, as equation (18.9) indicates. Therefore, degrees of severity share the problems of p-values and are poor indicators of rational credibility, at least from a Bayesian point of view.

Mayo might counter that a result by Casella and Berger (1987) shows the convergence of Bayesian and frequentist measures of evidence in the one-sided testing problem, effectively alleviating the Bayesian's worries. But I am skeptical that this response works. The reconciliationist results by Casella and Berger make substantial demands on the statistical model, e.g., probability density functions must be symmetric with monotone likelihood ratios (e.g., Theorem 3.1). Even then, the p-value can still substantially deviate from the posterior probability. Only for very large datasets, we will finally have agreement between Bayesian and frequentist measures.

Thirdly, Mayo only provides an evidential interpretation of *directional* tests, not a rebuttal of the objections raised against two-sided frequentist tests (e.g., in Lindley's Paradox). In particular, the question of whether a specific model can be treated as a proxy for a more general model is not addressed in the error-statistical framework: it specifies only warranted differences from a point hypothesis in a particular direction. However, a statistical framework that aims at resolving the conceptual problems in frequentist inference should address these concerns, too.

Fourthly, as I will argue in the following section, the error-statistical theory fails, like any frequentist approach, to give a satisfactory treatment of the optional stopping problem.

18.7 Sequential Analysis and Optional Stopping

Sequential analysis is a form of experimental design where the sample size is not fixed in advance. This is of great importance in clinical trials, e.g., when we test the efficacy of a

medical drug and compare it to the results in a control group. In those trials, continuation of the trial (and possibly the decision to allocate a patient to either group) depends on the data collected so far. For instance, data monitoring committees will decide to stop the trial as soon as there are substantial signs that the tested drug has harmful side effects.

A **stopping rule** describes under which conditions a sequential trial is terminated, as a function of the observed results. For example, we may terminate a trial when a certain sample size is reached, or whenever the results clearly favor one of the two tested hypotheses.[12] The dissent between Bayesians and frequentists concerns the question of whether our inference about the efficacy of the drug should be sensitive to the specific stopping rule used.

From a frequentist point of view, the significance of a result may depend on whether or not it has been generated by a fixed sample-size experiment. Therefore, regulatory bodies such as the Food and Drug Administration (FDA) require experimenters to publish all trial properties in advance, *including the stopping rule they are going to use*.

For a Bayesian, the LP implies that only information contained in the likelihood function affects a post-experimental inference. Since the likelihood functions of the parameter values under different stopping rules are proportional to each other (proof omitted), stopping rules can have no evidential role. Berger and Berry (1988, p. 34) call this the **Stopping Rule Principle**. To motivate this principle, Bayesians argue that

> The design of a sequential experiment is .. what the experimenter actually *intended* to do. (Savage 1962, p. 76. Cf. Edwards, Lindman, and Savage (1963, p. 239).)

In other words, since such intentions are "locked up in [the experimenter's] head" [ibid.], not verifiable for others, and apparently not causally linked to the data-generating process, they should not matter for sound statistical inference. This is the **sample space dependence** of frequentist inference mentioned in Section 18.4.

This position has substantial practical advantages: if trials are terminated for unforeseen reasons, e.g., because funds are exhausted or because unexpected side effects occur, the observed data can be interpreted properly in a Bayesian framework, but not in a frequentist framework. As externally forced discontinuations of sequential trials frequently happen in practice, claims for the evidential relevance of stopping rules would severely compromise the proper interpretation of sequential trials.

However, from a frequentist point of view, certain stopping rules, such as sampling on until the result favors a particular hypothesis, lead us to biased conclusions (cf. Mayo 1996, pp. 343–45). In other words, neglect of stopping rules in the evaluation of an experiment makes us *reason to a foregone conclusion*. Consider a stopping rule that rejects a point null $H_0 : \theta = \theta_0$ in favor of $H_0 : \theta \neq \theta_0$ whenever the data are significant at the 5% level. With probability one, this event will happen at *some* point, independent of the true value of θ (Savage 1962; Mayo and Kruse 2001).[13] In this case, the type I error is apparently 0.05 while it actually approaches unity, since rejection of the null is bound to happen at some

[12] Formally, stopping rules are functions $\tau : (\mathcal{X}^\infty, \mathcal{A}^\infty) \to \mathbb{N}$ from the measurable space $(\mathcal{X}^\infty, \mathcal{A}^\infty)$ (=the infinite product of the sample space) to the natural numbers such that for each $n \in \mathbb{N}$, the set $\{x \in \mathcal{X}^\infty | \tau(x) = n\}$ is measurable.

[13] As we saw in the case of Lindley's Paradox, an ever smaller divergence from the null is sufficient to trigger statistical significance as sample size increases.

point. Not only this: a malicious scientist who wants to publish a result where a certain null hypothesis is rejected can design an experiment where this will almost certainly happen, with an arbitrarily high level of statistical significance (provided she does not run out of money before). Should we trust the scientist's conclusion? Apparently no, but the Bayesian cannot tell why. Frequentists such as Mayo read this as the fatal blow for positions that deny the post-experimental relevance of stopping rules.

The Bayesian response is threefold. First, the posterior probability of a hypothesis cannot be arbitrarily manipulated (Kadane et al. 1996). If we stop an experiment if and only if the posterior of a hypothesis passes a certain threshold, there will be a substantial chance that the experiment never terminates. It is therefore not possible to reason to a foregone conclusion with certainty by choosing a suitable stopping rule. Similar results that bound the probability of observing misleading evidence have been proved by Savage (1962) and Royall (2000).

Secondly, the frequentist argument is valid only if frequentist evidence standards are assumed. But from a Bayesian point of view, even biased experiments can produce impressive evidence—provided the design of the experiment did not interfere with the data-generating mechanism. If scientists had to throw away arduously collected data just because the experimental design was not properly controlled, scientific knowledge would not be at the point where it is now.

Thirdly, preferring a (post-experimental) decision rule that is sensitive to the used stopping rule leads to incoherence, in the sense that a Dutch Book can be construed against such preferences. This result by Sprenger (2009) is derived from a more general, quite technical theorem by Kadane, Schervish and Seidenfeld (2003).

These arguments demonstrate the coherence of the Bayesian approach to stopping rules, and show that they should not matter post-experimentally if statistics is supposed to be consistent with standard theories of rational preferences and decisions. That said, there is a valid core in the frequentist argument: sequential trials are often costly and require careful pre-experimental design for efficient experimentation. Also, the termination of a sequential trial often involves complex ethical issues. Here, the choice of a stopping rule can make a great difference to frequentists *and* Bayesians.

18.8 Discussion: Some Thoughts on Objectivity

We have introduced the Bayesian and the frequentist paradigms as well as their philosophical foundations, and focused on three grand questions: what should we believe, what should we do, and how should we measure statistical evidence? In particular the last question sparks fierce debates between Bayesians and frequentists, as well as between different strands of frequentism.

The author has not concealed his inclinations toward a broadly Bayesian view on inductive inference. This position is supported by the numerous inadequacies of significance tests and p-values, among which are the mathematical incompatibility with posterior probabilities, the neglect of effect size, and Lindley's Paradox. Moreover, the frequentist stance on stopping rules appears to lead to unacceptable consequences.

In light of these arguments, it may be surprising that frequentist statistics is still the dominating school of inductive inference in science. However, two points have to be considered. First, there are still principled reservations against the subjectivist approach because it apparently threatens the objectivity, impartiality and epistemic authority of science. Although the ideal of **objective statistical inference** as free from personal perspective has been heavily criticized (e.g., Douglas 2009) and may have lost its appeal for many philosophers, it is still influential for many scientists and regulatory agencies who are afraid of external interests influencing the inference process. For a long time, bodies such as the FDA were afraid that Bayesian analysis would be misused for discarding hard scientific evidence on the basis of prejudiced a priori attitudes, and only recently, the FDA has opened up to a Bayesian analysis of clinical trials.

Secondly, scientific institutions such as editorial offices, regulatory bodies and professional associations are inert: they tend to stick to practices which have been "well probed" and with which they are familiar. Take experimental psychology as an example: even implementing the most basic changes, such as accompanying p-values by effect size estimates and/or power calculations, was a cumbersome process that took a lot of time. Changing the relevant textbook literature and the education of young scientists may take even more time. On a positive note, a more pluralist climate has developed over recent years, and there is now an increasing interest in Bayesian and other non-orthodox statistical methods.

Thirdly, even some well-known Bayesian modelers such as Gelman and Shalizi (2013) confess that while they apply Bayesian statistics as a technical tool, they would not qualify themselves as subjectivists. Their methodological approach stands in the hypothetico-deductive tradition: constructing a model, deriving predictions from the model, and assessing the model on the basis of its success. This is similar to the frequentist rationale of hypothesis testing. While Bayesians may have the winning hand from a purely foundational point of view, it is by no means obvious that their methods provide the best answer in scientific practice. This points us to the task of telling a story of how Bayesian inference relates to statistical model checking in a hypothetico-deductive spirit, and more generally, to investigating the relationship between qualitative and quantitative, between subjective and objective accounts of theory confirmation (Sprenger 2013b).

Finally, I would like to mention some compromises between Bayesian and frequentist inference that Bayesians have invented for meeting objectivity demands. First, there is the **conditional frequentist approach** of Berger (2003) and his collaborators (e.g., Berger, Brown and Wolpert 1994). The idea of this approach is to supplement frequentist inference by conditioning on the observed strength of the evidence (e.g., the value of the Bayes factor). The resulting hypothesis tests have a valid interpretation from a Bayesian and a frequentist perspective and are therefore acceptable for either camp. Nardini and Sprenger (2013) describe how this approach can ameliorate the practice on sequential trials in medicine. Secondly, there are José Bernardo's (2012) **reference priors**, which are motivated by the idea of maximizing the information in the data vis-à-vis the prior and posterior distribution (see Sprenger 2012, for a philosophical discussion).

Attempts to find a compromise between Bayesian and frequentist inference are, for the most part, still terra incognita from a philosophical point of view. In my perspective, there is a lot to gain from carefully studying how these approaches try to find a middle ground between subjective Bayesianism and frequentism.

References

Aldrich, J. (2016) The Origins of Modern Statistics. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Berger, J. O. (2003) Could Fisher, Jeffreys and Neyman Have Agreed on Testing? *Statistical Science.* 18. pp. 1–32.

Berger, J. O. and Berry, D. (1988) The Relevance of Stopping Rules in Statistical Inference (with discussion). In Gupta, S. and Berger, J. O. (eds.) *Statistical Decision Theory and Related Topics IV*. 29–72. New York, NY: Springer.

Berger, J. O., Brown, L. D., and Wolpert, R. L. (1994) A Unified Conditional Frequentist and Bayesian Test for Fixed and Sequential Simple Hypothesis Testing. *Annals of Statistics.* 22. pp. 1787–807.

Berger, J. O. and Delampady, M. (1987) Testing Precise Hypotheses. *Statistical Science.* 2. pp. 317–52.

Berger, J. O. and Sellke, T. (1987) Testing a Point Null Hypothesis: The Irreconciliability of P-values and Evidence. *Journal of the American Statistical Association.* 82. pp. 112–39.

Berger, J. O. and Wolpert, R. L. (1984) *The Likelihood Principle*. Hayward/CA: Institute of Mathematical Statistics.

Bernardo, J. M. (1999) Nested Hypothesis Testing: The Bayesian Reference Criterion. In Bernardo, J. et al. (eds.) *Bayesian Statistics 6: Proceedings of the Sixth Valencia Meeting*. pp. 101–30. Oxford: Oxford University Press.

Bernardo, J. M. (2012) Integrated Objective Bayesian Estimation and Hypothesis Testing. In Bernardo, J. M. et al. (eds.) *Bayesian Statistics 9: Proceedings of the Ninth Valencia Meeting*. pp. 1–68. Oxford: Oxford University Press.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Birnbaum, A. (1962) On the Foundations of Statistical Inference. *Journal of the American Statistical Association.* 57. pp. 269–306.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago, IL: The University of Chicago Press.

Casella, G. and Berger, R. L. (1987) Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem. *Journal of the American Statistical Association*. 82. 106–11.

Cohen, J. (1994) The Earth is Round ($p < .05$). *American Psychologist.* 49. pp. 997–1001.

Cumming, G. and Finch, S. (2005) Inference by Eye: Confidence Intervals, and How to Read Pictures of Data. *American Psychologist.* 60. pp. 170–80.

Douglas, H. (2009) *Science, Policy and the Value-Free Ideal*. Pittsburgh, PA: University of Pittsburgh Press.

Edwards, A. W. F. (1972) *Likelihood*. Cambridge: Cambridge University Press.

Edwards, W., Lindman, H., and Savage, L. J. (1963) Bayesian Statistical Inference for Psychological Research. *Psychological Review.* 70. pp. 450–99.

Fidler, F. (2005) *From Statistical Significance to Effect Estimation*. Ph.D. Thesis: University of Melbourne.

Fisher, R. A. (1925) *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.

Fisher, R. A. (1956) *Statistical Methods and Scientific Inference*. New York, NY: Hafner.

Gelman, A. and Shalizi, C. (2013) Philosophy and the Practice of Bayesian Statistics (with discussion). *British Journal of Mathematical and Statistical Psychology.* 66. pp. 8–18.

Gigerenzer, Gerd (1993) The Superego, the Ego, and the Id in Statistical Reasoning. In Keren, G. and Lewis, C. (eds.) *A Handbook for Data Analysis in the Behavioral Sciences.* pp. 311–39. Hillsdale, NJ: Erlbaum.

Gillies, D. (1971) A Falsifying Rule for Probability Statements. *British Journal for the Philosophy of Science.* 22. pp. 231–61.

Goodman, S. N. (1999) Towards Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine.* 130. pp. 1005–13.

Hacking, Ian (1965) *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Harlow, L. L., Mulaik, S. A. and Steiger, J. H. (eds.) (1997) *What If There Were No Significance Tests?* Mahwah, NJ: Erlbaum.

Hartmann, S. and Sprenger, J. (2011) Mathematics and Statistics in the Social Sciences. In Jarvie, I. C. and Bonilla, J. Z. (eds.) *SAGE Handbook of the Philosophy of Social Sciences*. 594–612. London: SAGE.

Hoover, K. D. and Siegler, M. V. (2008) The Rhetoric of 'Signifying Nothing': A Rejoinder to Ziliak and McCloskey. *Journal of Economic Methodology.* 15. pp. 57–68.

Howson, C. and Urbach, P. (2006) *Scientific Reasoning: The Bayesian approach*. 3rd ed. La Salle: Open Court.

Ioannides, J. P. A. (2005). Why Most Published Research Findings Are False. *PLOS Medicine* 2.8. e124. [Online] Available from doi:10.1371/journal.pmed.0020124 [Accessed 30 Sep 2015].

Jeffreys, H. (1939) *Theory of Probability*. Oxford: Clarendon Press.

Kadane, J. B., Schervish, M. J., and Seidenfeld, T. (1996) When Several Bayesians Agree That There Will Be No Reasoning to a Foregone Conclusion. *Philosophy of Science.* 63. pp. S281–S289.

Kass, R. and Raftery, A. (1995) Bayes Factors. *Journal of the American Statistical Association.* 90. pp. 773–90.

Krüger, L., Gigerenzer, G., and Morgan, M. (eds.) (1987) *The Probabilistic Revolution*, Vol. 2: *Ideas in the Sciences*. Cambridge, MA: The MIT Press.

La Caze, A. (2016) Frequentism. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Lele, S. (2004) Evidence Functions and the Optimality of the Law of Likelihood (with discussion). In Taper, M. and Lele, S. (eds.) *The Nature of Scientific Evidence*. pp. 191–216. Chicago, IL and London: The University of Chicago Press.

Lindley, D. V. (1957) A Statistical Paradox. *Biometrika.* 44. pp. 187–92.

Mayo, D. G. (1996) *Error and the Growth of Experimental Knowledge*. Chicago, IL and London: The University of Chicago Press.

Mayo, D. G. (2010) An Error in the Argument from Conditionality and Sufficiency to the Likelihood Principle. In Mayo, D., Spanos, A. (eds.) *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability, and the Objectivity and Rationality of Science*. pp. 305–14. Cambridge: Cambridge University Press.

Mayo, D. G. and Kruse, M. (2001) Principles of Inference and Their Consequences. In Cornfield, D. and Williamson, J. (eds.) *Foundations of Bayesianism*. pp. 381–403. Dordrecht: Kluwer.

Mayo, D. G. and Spanos, A. (2006) Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction. *The British Journal for the Philosophy of Science.* 57. pp. 323–57.

McCloskey, D. N. and Ziliak, S. T. (1996) The Standard Error of Regressions. *Journal of Economic Literature.* 34. pp. 97–114.

Nardini, C. and Sprenger, J. (2013) Bias and Conditioning in Sequential Medical Trials. *Philosophy of Science*. 80. pp. 1053–64.

Neyman, J. and Pearson, E. (1933) On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society A*. 231. pp. 289–337.

Neyman, J. and Pearson, E. (1967) *Joint Statistical Papers*. Cambridge: Cambridge University Press.

Oakes, M. (1986) *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York, NY: Wiley.

O'Hagan, T. (2012) Posting on the statistical methods used in the discovery of the Higgs Boson, made via the email list of the International Society for Bayesian Analysis (ISBA). Available from: www.isba.org [Accessed 30 Sep 2015].

Popper, K. R. (1934/1959) *Logic of Scientific Discovery*. Translated from the German by the author. London: Hutchinson. (Originally published in German).

Romeijn, J. W. (2010) Inductive Logic and Statistics. In Gabbay, D., Hartmann, S., and Woods, J. (eds.) *Handbook of the History of Logic.* Vol. 10 *(Inductive Logic)*. 625–50. Amsterdam: Elsevier.

Royall, R. (1997) *Scientific Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Royall, R. (2000) On the Probability of Observing Misleading Statistical Evidence. *Journal of the American Statistical Association.* 95. pp. 760–68.

Savage, L. J. (1962) *The Foundations of Statistical Inference*. London: Methuen.

Spanos, A. (2010) Is Frequentist Testing Vulnerable to the Base-Rate Fallacy? *Philosophy of Science.* 77. pp. 565–83.

Spielman, S. (1974) The Logic of Significance Testing. *Philosophy of Science.* 41. pp. 211–25.

Spielman, S. (1978) Statistical Dogma and the Logic of Significance Testing. *Philosophy of Science.* 45. pp. 120–35.

Sprenger, J. (2009) Evidence and Experimental Design in Sequential Trials. *Philosophy of Science.* 76. pp. 637–49.

Sprenger, J. (2012) The Renegade Subjectivist: Jose Bernardo's Reference Bayesianism. *Rationality, Markets and Morality.* 3. pp. 1–13.

Sprenger, J. (2013a) Testing a Precise Null Hypothesis: The Case of Lindley's Paradox. *Philosophy of Science*. 80. pp. 733–44.

Sprenger, J. (2013b) A Synthesis of Hempelian and Hypothetico-Deductive Confirmation. *Erkenntnis*. 78. pp. 727–38.

Williamson, J. (2010) *In Defense of Objective Bayesianism*. Oxford: Oxford University Press.

Ziliak, S. T. and McCloskey, D. N. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice and Lives*. Ann Arbor, MI: University of Michigan Press.

Zynda, L. (2016) Subjectivism. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

CHAPTER 19

THE PROPENSITY INTERPRETATION

DONALD GILLIES

19.1 Popper's Introduction of the Propensity Interpretation

The propensity interpretation of probability, or propensity theory of probability, was introduced by Popper in 1957[1], and subsequently expounded and developed by him in a series of papers and books (1959, 1967, 1983, 1990). As we shall see, the basic ideas of the propensity theory were subsequently taken up by other philosophers of science, who produced versions of the theory that often differed from Popper's. However, it seems natural to begin by considering Popper's views.

In his *Logic of Scientific Discovery* (1934/1972: pp. 146–214), Popper advocated a version of the frequency theory of probability, which was a modification of the theory that von Mises had presented in his 1928. Further reflection, however, gradually convinced Popper that the frequency theory was inadequate, and this led him to introduce the propensity theory. As the propensity theory was thus developed in explicit contrast to the frequency theory in a version similar to that of von Mises, it will be helpful to say a few things about this frequency theory before we go on to consider the propensity theory.

Von Mises regarded probabilities as objective in the sense that they were features of the natural or social worlds that had nothing to do with human beliefs or knowledge. Probability theory, according to von Mises, is a mathematical science dealing with objectively occurring random phenomena. Probability itself is an objective concept like mass in theoretical mechanics, or charge in electromagnetic theory. In the natural or social world, there are what von Mises calls 'collectives' where 'either the same event repeats itself again and again, or a great number of uniform elements are involved at the same time' (1928/1961: p. 11). The probability of an attribute in a collective is defined as its limiting frequency.

[1] Popper first presented the propensity theory at a conference in the University of Bristol. However, as he could not attend himself, his paper (1957) was read by his then student Paul K. Feyerabend.

Probabilities, then, are associated with collectives and are considered to be objective and independent of the individual who estimates them, just as the masses of bodies in mechanics are independent of the person who measures them.

Von Mises' emphasis on the objectivity of probabilities was something that appealed strongly to Popper, and he retained this feature of von Mises' theory when he changed to the propensity theory. However, there was another aspect of von Mises' theory which, even in 1934, Popper found unsatisfactory. This was von Mises' discussion of probabilities for single events, or *singular probabilities*. Von Mises denied that such probabilities could validly be introduced. The example he considered was the probability of death. We can certainly introduce the probability of death before 41 in a sequence of, say, 40-year-old Englishmen. It is simply the limiting frequency of those in the sequence who die before 41. But can we also introduce the probability of death before 41 for a particular 40-year old Englishman (Mr Smith, say)? Von Mises answered, 'no!' (1928/1961: p. 11):

> We can say nothing about the probability of death of an individual even if we know his condition of life and health in detail. The phrase "probability of death", when it refers to a single person has no meaning at all for us. This is one of the most important consequences of our definition of probability . . .

Of course it is easy to introduce singular probabilities on a version of the subjective theory which interprets probabilities as the degrees of belief of individuals measured by the rates at which they are prepared to bet. All Mr Smith's friends could, for example, take bets on his dying before 41, and hence introduce subjective probabilities for this event. Clearly, however, this procedure would not satisfy an objectivist like Popper. The key question for him was whether it was possible to introduce objective probabilities for single events.

Already in 1934 Popper disagreed with von Mises' denial of the possibility of objective singular probabilities. This was partly because he wanted such probabilities for his interpretation of quantum mechanics. Consider, for example, the two-slit experiment. This can be performed in such a way that only one photon is present in the apparatus at any time. Yet the diffraction pattern appears just as it does when photons pass through the apparatus *en masse*. Thus, so it could be argued, the probabilistic laws of quantum mechanics must apply to each single photon just as much as to a collective of photons. In 1934, Popper thought that objective singular probabilities could be introduced by a simple modification of von Mises' theory. Let us consider a single event that is a member of one of von Mises' collectives. We can then define its singular probability as the same as the probability of the event in the collective as a whole. Later on, however, in his 1957 and 1959, Popper presented an objection, which he himself had invented, to this way of defining singular probabilities, and this led him to his new theory of probability.

Popper's argument is as follows. Begin by considering two dice: one regular, and the other biased so that the probability of getting a particular face (say the 5) is 1/4. Now consider a sequence consisting almost entirely of throws of the biased die but with one or two throws of the regular die interspersed. Let us take one of these interspersed throws and ask what is the probability of getting a 5 on that throw. According to Popper's 1934 suggestion, this probability must be 1/4, because the throw is part of a collective for which prob(5) = 1/4. But this is an intuitive paradox, since it is surely much more reasonable to say that prob(5) = 1/6 for any throw of the regular die.

One way out of the difficulty is to modify the concept of collective so that the sequence of throws of the biased die with some throws of the regular die interspersed is not a genuine collective. The problem then disappears. This is just what Popper did (1959: p. 34):

> All this means that the frequency theorist is forced to introduce a modification of his theory – apparently a very slight one. He will now say that an admissible sequence of events (a reference sequence, a "collective") must always be a sequence of repeated experiments. Or more generally, he will say that admissible sequences must be either virtual or actual sequences which are *characterised by a set of generating conditions* – by a set of conditions whose repeated realisation produces the elements of the sequence.

He then continued a few lines later

> Yet, if we look more closely at this apparently slight modification, then we find that it amounts to a transition from the frequency interpretation to the propensity interpretation.

In this interpretation, the generating conditions are considered as endowed with a propensity to produce the observed frequencies. As Popper put it (1959: p. 35):

> But this means that we have to visualise the conditions as endowed with a tendency or disposition, or propensity, to produce sequences whose frequencies are equal to the probabilities; which is precisely what the propensity interpretation asserts.

And, in similar fashion (1957: p. 67):

> . . . since the probabilities turn out to depend upon the experimental arrangement, they may be looked upon as *properties of this arrangement. They characterize the disposition, or the propensity*, of the experimental arrangement to give rise to certain characteristic frequencies *when the experiment is often repeated.*

So Popper's notion of propensity involves the change from collectives to conditions, but there is more to it than that. The word 'propensity' suggests some kind of dispositional account, and this marks another difference from the frequency view. A useful way of looking into this matter will be to consider some earlier views of Peirce's which were along the same lines.[2] These are contained in the following passage (1910/1957: pp. 79–80):

> I am, then, to define the meaning of the statement that the *probability*, that if a die be thrown from a dice box it will turn up a number divisible by three, is one-third. The statement means that the die has a certain 'would-be'; and to say that the die has a 'would-be' is to say that it has a property, quite analogous to any *habit* that a man might have. Only the 'would-be' of the die is presumably as much simpler and more definite than the man's habit as the die's homogeneous composition and cubical shape is simpler than the nature of the man's nervous system and soul; and just as it would be necessary, in order to define a man's habit, to describe how it would lead him to behave and upon what sort of occasion – albeit this statement would by no means imply that the habit *consists* in that action – so to define the die's 'would-be' it is necessary to say how it would lead the die to behave on an occasion that would bring out the full consequence of the 'would-be'; and this statement will not of itself imply that the 'would-be' of the die *consists* in such behavior.

Peirce then goes on to describe an occasion that would bring out the full consequence of the 'would-be'. Such an occasion is an endless sequence of throws of the die and the relevant

[2] On this topic see Fetzer (1993).

behaviour of the die is that the appropriate relative frequencies fluctuate round the value of 1/3. Though Peirce does not actually say this, he could have added that the relative frequencies gradually come closer and closer to 1/3, and eventually converge on it.

Peirce's 'would-be' is obviously similar to Popper's 'propensity', but there is an interesting difference. Peirce regards his 'would-be' as a property of the die, whereas Popper regards his 'propensity' as a property not just of the die but also of the conditions under which it is thrown. Mellor in 1971 develops a version of the propensity theory that sides with Peirce rather than Popper on this point. However, Popper has produced a number of interesting examples to support his position. One of these concerns tossing an ordinary coin, for which the probability of heads is 1/2. We can alter the conditions of tossing by letting the coin fall not on a flat surface such as a table top, but on a surface in which a large number of slots have been cut. We now no longer have two outcomes, 'heads' and 'tails' but *three*, viz. 'heads', 'tails', and 'edge', the third outcome being that the coin sticks in one of the slots. Further, because 'edge' will have a positive probability, the probabilities of both 'heads' and 'tails' will be reduced. This example shows that not only do the probabilities of outcomes change with the manner of tossing but even that the set of outcomes can similarly vary.

Despite this difference between Peirce and Popper, the two agree on holding that probabilities are dispositions, and Peirce makes a valuable distinction between probability as a dispositional quantity, and an occasion which would bring out the full consequence of this disposition. The importance of making this distinction is that it allows us to introduce probabilities as dispositions even on occasions where the full consequence of the disposition is not manifested, where in effect we do not have a long sequence of repetitions.

This shows the difference between a dispositional account of probability and von Mises' frequency theory, for von Mises held that probabilities ought only to be introduced in physical situations where we have an empirical collective, i.e. a long sequence of events. If we adopt Popper's propensity theory, however, it becomes perfectly legitimate to introduce probabilities on a set of conditions *even though these conditions are not repeated a large number of times*. We are allowed to postulate probabilities (and might even obtain testable consequences of such a postulation) when the relevant conditions are repeated only once or twice. Thus Popper's propensity theory provides a valuable extension of the situations to which probability theory applies, as compared to von Mises' frequency view.

So far then, we have found 3 ingredients in Popper's early propensity theory. These are: (1) probabilities are associated with repeatable conditions rather than with collectives; (2) probabilities are dispositions; and (3) objective probabilities can be introduced for single events. Now (1) and (2) are relatively unproblematic, but (3) is a different matter. Several philosophers of science have argued that there cannot be objective probabilities for single events. In the next section, I will examine some of these arguments, and discuss what bearing they have on the propensity interpretation of probability.

19.2 Can There be Objective Probabilities of Single Events?

Ayer (1963: pp. 198–208) discussed what is perhaps the major difficulty in the way of introducing objective probabilities for single events, though the problem has an earlier

history. The difficulty is this. Suppose we are trying to assign a probability to a particular event. The probability will vary according to the set of conditions which the event is considered as instantiating – according, in effect, to how we describe the event. But then we are forced to consider the probabilities as attached to the conditions that describe the event rather than to the event itself.

To illustrate this, let us return to our example of the probability of a particular man aged 40 living to be 41. The probability would seem to vary depending on whether we regard the individual merely as a man or more particularly as an Englishman; for the life expectancy of Englishmen is higher than that of mankind as a whole. Similarly the probability will alter depending on whether we regard the individual as an Englishman aged 40 or as an Englishman aged 40 who smokes two packets of cigarettes a day, and so on. This does seem to show that probabilities should be considered as dependent on the properties used to characterize an event rather than as dependent on the event itself.

It is natural in the context of the propensity theory to consider the problem in terms of the conditions used to describe a particular event, but we could equally well look at the problem as being that of assigning the event to a reference class. Instead of asking whether we should regard Mr Smith as a man aged 40, or as an Englishman aged 40, or as an Englishman aged 40 who smokes two packets of cigarettes a day, we could ask equivalently whether we should assign him to the reference class of all men aged 40, of all Englishmen aged 40, or of all Englishmen aged 40 who smoke two packets of cigarettes a day. The reference-class formulation is more natural in the context of the frequency theory where the problem first appeared. Although I am discussing the propensity theory, I will continue to use the traditional terminology, and refer to this fundamental problem as *the reference-class problem*.

Howson and Urbach argued in 1989 that the reference-class problem shows that single-case probabilities are subjective rather than objective. However, they also suggested that singular probabilities, though subjective, may be based on objective probabilities. Suppose, for example, that the only relevant information that Mr B has about Mr A is that Mr A is a 40-year-old Englishman. Suppose Mr B has a good estimate (p, say) of the objective probability of 40-year-old Englishmen living to be 41. Then it would be reasonable for Mr B to put his subjective betting quotient on Mr A's living to be 41 equal to p, and thereby make his subjective probability objectively based. This does not, however, turn Mr B's subjective probability into an objective one, for consider Mr C, who knows that Mr A smokes two packets of cigarettes a day, and who also has a good estimate of the objective probability (q, say) of 40-year-old Englishmen who smoke two packets of cigarettes a day living to be 41. Mr C will put his subjective probability on the same event (Mr A living to be 41) at a value q different from Mr B's value p. Once again the probability depends on how the event is categorized rather than on the event itself. Howson and Urbach put the point as follows (1989: p. 228):

> . . . single-case probabilities . . . are not themselves objective. They are subjective probabilities, which considerations of consistency nevertheless dictate must be set equal to the objective probabilities just when all you know about the single case is that it is an instance of the relevant collective. Now this is in fact all that anybody ever wanted from a theory of single-case probabilities: they were to be equal to objective probabilities in just those conditions. The incoherent doctrine of objective single-case probabilities arose simply because people failed to mark the subtle distinction between the values of a probability being objectively based and the probability itself being an objective probability.

Howson and Urbach's argument against the existence of objective probabilities for single events is certainly a strong one. There are, however, some arguments on the other side in favour of objective singular probabilities. Here is one.[3] It might be conceded that it is difficult to assign such singular probabilities to events involving humans, such as Mr Smith, our 40-year-old Englishman, dying before he is 41, or having a car accident before he is 41. However, it could still be claimed that such singular probabilities are more plausible in cases like games of chance, or scientific experiments such as the quantum-mechanical two-slit experiment with electrons. In such cases, it could be argued, we have a well-defined reference class consisting of a sequence of tosses of a coin, rolls of a die, or experiments, which are carried out under a specified set of repeatable conditions. Here it seems reasonable to say that we should assign an objective singular probability on each toss, roll, or experiment equal to the objective probability in the reference class as a whole.

In the light of this, we can say that the introduction of objective probabilities for single events is rather problematic. There are some arguments in favour, but also strong arguments, based on the reference-class problem, against. Now suppose someone decides that objective probabilities for single events are illegitimate. Does that mean that he or she has given up the propensity theory of probability? If we identify the propensity theory of probability with the position that Popper defended in his 1957 and 1959 papers, then the answer must be 'yes', for Popper in those works regards the introduction of objective singular probabilities as the most important feature of his new theory. However, as we saw earlier, Popper's new propensity theory had other features which are not strongly connected with his advocacy of objective singular probabilities. For example, the new theory holds that probabilities be assigned to sets of repeatable conditions (**S**) rather than to collectives (**C**), and that probabilities be analysed as dispositions. This suggests that we might use the term 'propensity theory of probability' in a more general sense, in which it would be possible to deny the existence of objective singular probabilities while still supporting a version of the propensity theory.

Of course the question we are considering is a semantic one. Should we define 'the propensity theory of probability' as the view expounded by Popper in his 1959, or should we give this term a more general sense? There is an important consideration which favours the second alternative. Popper's suggestion of a new philosophical view of probability aroused a great deal of interest in the community of philosophers of science. As a result, many different philosophers of science developed different versions of the new propensity theory. I have already mentioned Mellor (1971) as an example of this, and, without aiming to be completely comprehensive, I will give further examples in what follows. Moreover, as Runde points out in his 1996, Popper's later views on propensity, particularly in his 1990, differ considerably from his earlier views. Thus Popper's 1959 paper stimulated the development of a number of different, but related, philosophical theories of probability. It seems best to use the term 'propensity theory of probability' as a general description of members of this group of theories.[4]

[3] This argument was suggested to me by Ladislav Kvasz.

[4] This is the way I use 'propensity theory' at present, but, in some of my earlier writings, I used a narrower definition of propensity. For example, in my *An Objective Theory of Probability* (1973), I argued that the theory developed in the book was not a propensity theory because it differed in some respects from Popper's 1959 theory. Subsequently, however, the term 'propensity' became well established in the

If, however, we do not have a single propensity theory of probability, but rather several different propensity theories of probability, the question arises of how these different propensity theories are to be classified. In the next section, I will propose a classification, and give some examples of the different kinds of propensity theory.

19.3 Classification of Propensity Theories

Propensity theories in the general sense just explained can be classified into (i) *single-case propensity theories*, and (ii) *long-run propensity theories*.[5] A single-case propensity theory is one in which propensities are regarded as propensities to produce a particular result on a specific occasion. Now Popper wanted his propensity theory to apply in the single case, and yet some of his formulations of the theory in his 1959 referred to long-run frequencies. This aspect of Popper's thinking about propensity is discussed by Giere, who writes (1973: p. 472):

> ... an early discussion of Popper's includes the statement that "propensities turn out to be *propensities to realize singular events*" [1959: p. 28, Popper's italics]. This parallels my view that propensities are tendencies to produce specific outcomes on particular trials. Yet in the same article Popper describes his propensity interpretation as asserting precisely that experimental conditions are "endowed with a ... propensity to produce *sequences* whose *frequencies* are equal to the probabilities" [1959: p. 35, my italics]. ... in his later writings Popper seems to make the production of *sequences* fundamental.

Writing in 1973, Giere uses the phrase 'Popper's later writings' to refer to Popper's 1967 article: Quantum Mechanics without 'the Observer', and writes as follows about this article (pp. 472–3):

> Propensities, he says are "*properties of the repeatable experimental arrangement*" [1967, p. 38, Popper's italics]. ... But then the description of the propensity must include a specification of what counts as a repetition of *the* experiment. Popper accepts this "*relativity to specification*" [1967, p. 39, Popper's italics], but seems not to realize that it destroys his earlier claim to have solved the problem of the single case. ... Popper fails to solve the problem of the single case for the old-fashioned reason that he provides no solution to the problem of the reference class.

Giere seems to be correct here. Popper's original propensity theory was, in a sense, *both* long-run *and* single-case, but the two halves did not fit together very well. As soon as we associate propensities with repeatable conditions, then, as Giere observes, the

literature, and has taken on a broader meaning. I would therefore now reclassify my 1973 position as one particular example of a propensity theory. I had some discussions with Popper on this point after my book had appeared. Interestingly Popper favoured using the term 'propensity' in a general sense rather than as specifically referring to his own views.

[5] The distinction between long-run and single-case propensity theories is taken from Fetzer (1988: pp. 123, 125–6). However, I am using the terminology in a slightly different sense from Fetzer. Fetzer takes the 'long-run' to refer to infinite sequences, while I am using 'long-run' to refer to long, but still finite, sequences of repetitions, as well as infinite sequences.

reference-class problem arises, since any single event is an instance of many different sets of repeatable conditions. It is thus not clear that we can introduce propensities for single events. Giere's way out of the difficulty is to drop the long-run aspects of Popper's theory, and propose a purely single-case propensity theory. Despite his criticism of Popper, he follows Popper in regarding the probabilities in quantum mechanics as fundamental to the propensity theory, and indeed says (Giere 1973: p. 476) that he is 'taking quantum phenomena as paradigm examples of propensities.'

Another difficulty for Popper's long-run formulation of the propensity theory was pointed out by Hájek, who writes (2002/2011: section 3.5):

> According to the long-run theories, propensities are tendencies to produce relative frequencies with particular values, but the propensities are not probability values themselves; . . . . According to Popper, for example, a fair die has a propensity – an *extremely strong* tendency – to land "3" with a long-run frequency of 1/6. The small value of 1/6 does *not* measure this tendency.

Hájek has here raised an important objection to the long-run formulation of the propensity theory which Popper gave in 1959. As will be remembered, this ran (Popper 1959: p. 35):

> . . . we have to visualise the conditions as endowed with a tendency, or disposition, or propensity, to produce sequences whose frequencies are equal to the probabilities; which is precisely what the propensity theory asserts.

However, as Hájek points out, the propensity of the conditions of throwing a fair die to produce a long-run frequency of 3's (approximately) equal to 1/6 is very high, and definitely not equal to 1/6. Thus, if propensities are characterized as in the above quotation, they are not equal to probabilities, and so do not give an interpretation of the probability calculus – contrary to Popper's intention. The question arises as to whether we can reformulate the long-run propensity view so that it does become an interpretation of the probability calculus as originally intended. I will now attempt to give such a reformulation.[6]

Let us suppose that **S** is a set of repeatable conditions, and A is one of the outcomes of **S**. To say that the probability of A given **S** is p [$P(A|\mathbf{S}) = p$] is to claim that the conditions **S** have a propensity equal to p to produce A on any repetition of **S**.

It might be objected that this is not a long-run, but a single-case propensity theory since it speaks of the propensity to produce A on any repetition of **S**. A particular repetition of **S** is of course a single event. My answer to this difficulty is that a particular single event E, such as: A occurs to a particular individual at such and such time and place, may be the repetition of a number of different repeatable conditions $\mathbf{S_1}$, $\mathbf{S_2}$, . . . , and the propensities $P(A|\mathbf{S_1})$, $P(A|\mathbf{S_2})$. . . . may be different. This is illustrated by the standard example of Mr Smith, aged 40, dying before his 41^{st} birthday. Here A = dies before 41^{st} birthday, and the different sets of repeatable conditions are . . . is a 40-year-old man, . . . is a 40-year-old Englishman, . . . is a 40-year-old Englishman who smokes two packets of cigarettes a day . . . The particular event of Mr Smith's demise can be regarded as a repetition of any of these sets of repeatable conditions, and the propensity of A on the repetition is different for each

[6] In my earlier writings on the propensity, e.g. (2000a) and (2000b), I used a formulation of the long-run propensity theory similar to that given by Popper in his 1959. Alan Hájek pointed out to me the difficulty in this formulation, and, the following attempt at a reformulation is the result of discussions with him.

different set of repeatable conditions. Thus the propensity of A on a repetition does not give the propensity of the single event. In fact there may not be such a propensity.

Another way of looking at it would be to say that we are not, when considering a repetition of the conditions **S**, considering the single event in itself but only *qua* repetition of the conditions **S**. The propensity is associated not with the single event, but with these conditions. I would indeed take the association of propensities with repeatable conditions as characteristic of long-run propensity theories. There are a number of further difficulties with long-run propensity theories, but I will postpone the consideration of these until Section 19.5, and now turn instead to examining some ways in which the single-case propensity theory has been developed.

As we saw, Giere, writing in 1973, thought that Popper was veering in the direction of a long-run propensity theory. However, in his later 1990, Popper moves away from the long-run approach and emphasizes single-case propensity. A similar position is also developed by Miller in his 1994 and 1996. This position retains from the earlier Popper objective singular probabilities, but abandons the association of propensities with repeatable conditions. Instead propensities are associated with the whole physical situation, which I will take as referring to the state of the universe as a whole. Popper writes (1990: p. 17):

> . . . propensities in physics are properties of *the whole physical situation* and sometimes of the particular way in which a situation changes.

One reason for this change may have been the desire to preserve objective probabilities for single events. If propensities are associated with repeatable conditions, then, as we have seen, the reference-class problem calls into question whether we can extend these propensities to particular instances of these conditions. At all events, Miller is determined to retain objective singular probabilities. He writes (1994: p. 177):

> . . . the propensity interpretation . . . is an objectivist interpretation where single-case probabilities are supreme . . .

He goes on to say (1994: pp. 185–6):

> Propensities are not located in physical things, and not in local situations either. Strictly, every propensity (absolute or conditional) must be referred to the complete situation of the universe (or the light-cone) at the time. Propensities depend on the situation today, not on other situations, however similar. Only in this way do we attain the specificity required to resolve the problem of the single case.

The main problem with the 1990s views on propensity of Popper and Miller is that they appear to change the propensity theory from a scientific to a metaphysical theory. If propensities are ascribed to a set of repeatable conditions, then by repeating the conditions we can obtain frequencies that can be used to test the propensity assignment. If, on the other hand, propensities are ascribed to the 'complete situation of the universe . . . at the time', it is difficult, in view of the unique and unrepeatable character of this situation, to see how such propensity assignments could be tested. Miller seems to agree with this conclusion since he writes (1996: p. 139):

> The propensity interpretation of probability is inescapably metaphysical, not only because many propensities are postulated that are not open to empirical evaluation . . .

Popper writes in a similar vein (1990: p. 17):

> But in many kinds of events … the propensities cannot be measured because the relevant situation changes and cannot be repeated. This would hold, for example, for the different propensities of some of our evolutionary predecessors to give rise to chimpanzees and to ourselves. Propensities of this kind are, of course, not measurable, since the situation cannot be repeated. It is unique. Nevertheless, there is nothing to prevent us from supposing such propensities exist, and from estimating them speculatively.

Of course we can indeed estimate the propensities speculatively, but if these speculations cannot be tested against data, they are metaphysical in character.

Now there is nothing wrong with developing a metaphysical theory of propensities, and such a theory may be relevant to the discussion of old metaphysical questions such as the problem of free will and determinism. However, the propensity theory of probability could be developed with another aim, namely that of providing an interpretation of the probabilities that appear in such natural sciences as physics and biology. For a theory of this kind, probability assignments should be testable by empirical data, and this makes it desirable that they should be associated with repeatable conditions.

Fetzer's single-case propensity theory differs from that of Miller and the later Popper in that he does not associate propensities with the complete state of the universe. As Fetzer says (1982: p. 195):

> … it should not be thought that propensities for outcomes … depend, in general, upon the complete state of the world at a time rather than upon a complete set of (nomically and/or causally) relevant conditions … which happens to be instantiated in that world at that time.

In comparison with Miller and the later Popper, Fetzer takes a step closer to making propensities testable by empirical data, but some difficulties remain. If, as Fetzer suggests, we ascribe propensities to a complete set of (nomically and/or causally) relevant conditions, then in order to test a conjectured propensity value we must make a conjecture about the complete list of the conditions that are relevant. The required conjecture might often be difficult to formulate and hard to test, thereby rendering the propensities once again metaphysical rather than scientific.

There is more to be said about Fetzer's version of the propensity theory, but, before doing so, we must introduce a new theme. So far we have concentrated on the question of whether the propensity theory can provide an account of objective probabilities for single events. However, propensities have also been connected with causality. This aspect of the propensity theory is particular stressed by Popper in his 1990 book. Indeed the first part of this book (1990: pp. 1–26) is entitled 'A World of Propensities: Two New Views of Causality.' In the next section, I will examine the links between propensity and causality.

19.4 Propensity and Causality: Humphreys' Paradox

To relate propensity to causality, it is first necessary to distinguish between *deterministic* and *indeterministic* causality. 'A causes B' involves a deterministic notion of causality only

if, *ceteris paribus*, the instantiation of A is always followed by B. A simple example of deterministic causality is 'The sprinkler causes the grass to get wet.' Here 'The sprinkler' is short for 'The sprinkler being properly connected to a working water supply and turned on'; when that happens the grass always gets wet.

Deterministic causality is the traditional concept of causality which is analysed by 18th- and 19th-century philosophers such as Hume and Kant. Indeed Kant says in *The Critique of Pure Reason* (1787/1958: B5):

> . . . the very concept of a cause . . . manifestly contains the concept of a necessity of connection with an effect and of the strict universality of the rule . . .

So, according to Kant, if A causes B, and A occurs, then B is sure to follow. This is what we have called deterministic causality. In the 20th century, however, a new concept of causality emerged, in connection largely with medical epidemiology. The first example of this new, or indeterministic, type of causality was 'Smoking causes lung cancer'. This claim was first made around 1950, and was initially very controversial. Now, however, it has become a generally accepted causal law, and yet smoking is not always followed by lung cancer. In fact only about 5% of smokers get lung cancer.

Popper suggested in his 1990 that propensity might be a generalization of the notion of cause. As he puts it (1990: p. 20):

> Causation is just a special case of propensity: the case of propensity equal to 1 . . .

Using our terminology, Popper's point seems to be that indeterministic causes are propensities whose values are less than 1, and deterministic causes are propensities equal to 1. So to say that smoking causes lung cancer is just to say that smokers have a propensity to get lung cancer. In fact, this propensity is around 5%.

This thesis of Popper's is simple and attractive, but it gives rise to a difficulty, which has come to be known as *Humphreys' Paradox*. This can be stated as follows. Causes have a definite direction in time. So if A causes B and A occurs before B, then B does not cause A. Apart from a few speculations in theoretical physics, it is almost universally conceded that causes do not operate backwards in time. The situation is very different with probabilities. In general, if P(A|B) is defined, then so is P(B|A). Probabilities have a symmetry where causes are asymmetrical. It thus seems that propensity cannot after all be a generalization of cause.

This problem was first noticed by Humphreys, and first published by Salmon, with a reference to Humphreys, in 1979.

Here is a simple example which illustrates the difficulty. It is certainly true to say that rain causes mud, but false to say that mud causes rain. On the other hand, during a particular period (say a day) at a certain place and a certain time of year, it would be fairly easy to ascertain both P(mud|rain) and P(rain|mud). Both probabilities are well-defined, though only the first appears to be causal in character.

This problem was named 'Humphreys' paradox' by Fetzer (1981), and has given rise to much interesting discussion. A statement by Humphreys himself of the paradox is to be found in his 1985.

One attitude that could be taken to the paradox is simply to say that propensities are not causal in character. There may indeed be interesting relations between indeterministic causes and propensities, but the two concepts are different and should not be identified. Another possible attitude would be to say that propensities are always causal in character,

but that they cannot be identified with probabilities. This seems to be Fetzer's position for he writes (1982: p. 195):

> . . . by virtue of their "causal directedness", propensities cannot be properly formalized either as "absolute" or as "conditional" probabilities satisfying inverse as well as direct probability relations.

And again (1991: pp. 297–8):

> . . . that propensities are not probabilities (in the sense of satisfying standard axioms, such as Bayes's theorem) by virtue of their causal directedness was not generally recognized before the publication of Humphreys (1985).

On Fetzer's account then propensities do not satisfy the standard Kolmogorov axioms (see Kolmogorov (1933/1956)). Working with Nute, however, Fetzer developed an alternative set of axioms for propensities. This system, which he calls 'a probabilistic causal calculus' is presented in his (1981: pp. 59–67). It has the feature that

> . . . p may bring about q with the strength n (where p occurs prior to or simultaneous with q), whether or not q brings about p with *any* strength m.
>
> (Fetzer 1981: p. 284)

Fetzer, however, emphasizes that the Fetzer–Nute probabilistic causal calculus has many axioms which are definitely probabilistic in character. It might therefore be more accurate to describe the Fetzer–Nute calculus as a non-standard probability theory. As Fetzer himself says (1981: p. 285):

> Perhaps this means that the propensity construction has to be classified as a *non-standard* conception of probability, which does not preclude its importance *even as an interpretation of probability*! Non-Euclidean geometry first emerged as a *non-standard* conception of geometry, but its significance is none the less for that. Perhaps, therefore, the propensity construction of probability stands to standard accounts of probability just as non-Euclidean constructions of geometry stood to standard accounts of geometry before the advent of special and of general relativity.

Since 'non-standard' has connotations of 'non-standard analysis' it might be better to speak of the probabilistic causal calculus as a non-Kolmogorovian probability theory by analogy with non-Euclidean geometry.

The Fetzer–Nute suggestion of a non-Kolmogorovian probability theory is a bold and revolutionary one, but its revolutionary character will naturally create problems in its achieving general acceptance, since at present nearly all mathematicians accept the Kolmogorov axioms as the basis of probability theory. There is, moreover, an enormous body of theorems which have been proved from the Kolmogorov axioms. The mathematical community is unlikely to give up this formidable structure and substitute another for it unless there are very considerable gains in so doing. This is one reason for preferring a propensity theory which retains the standard Kolmogorov axioms, but gives up the claim that propensities have causal directedness. A long-run propensity theory of this type will be presented in the next section.[7]

[7] The treatment of the Humphreys' paradox in this section was rather concise. A more detailed treatment is to be found in my 2000a: pp. 825–33, or in my 2000b: pp. 129–36. These references give

19.5 Connecting Propensities to Frequencies

In Section 19.3, the following formulation of a long-run propensity theory was given. It was supposed that S is a set of repeatable conditions, and A is one of the outcomes of S. To say that the probability of A given S is p [P(A|S) = p] is to claim that the conditions S have a propensity equal to p to produce A on any repetition of S. Following the discussion in Section 19.4, it will further be assumed that propensities satisfy the Kolmogorov axioms, but not that they have causal directedness. This long-run propensity theory is thus a kind of minimal propensity theory, since it provides an interpretation of the standard calculus of probability, without claiming that there are objective singular probabilities or that propensities have a causal directedness. In this theory propensities can be considered as theoretical entities that are defined partly by their satisfaction of the Kolmogorov axioms. The problem of this section is how we connect this theoretical entity to observed frequencies.

When, in Section 19.3, we discussed the single-case propensity theories of Miller and the later Popper, and of Fetzer, we remarked that these theories were in danger of making propensity a metaphysical rather than a scientific concept. The same problem arises in the present long-run propensity theory. Suppose we postulate that the propensity of A, given S = p. How can we test this hypothesis h say against evidence, to see whether it is confirmed or refuted? We could repeat the underlying conditions S n times, and A might occur r times, so that the observed frequency of A is r/n. Now if r/n differs considerably from p, we might regard this as refuting our hypothesis, but not necessarily. We could say instead that n is not sufficiently large, and for a much larger number of repetitions, the relative frequency of A might come to be quite close to p. The problem here is to say what is meant by 'long' in the phrase 'a long series of repetitions'. There is also a problem about saying what is meant by 'approximately equal to p'. Suppose for example that p = 0.5, n = 2000, and that the observed frequency of A is 0.487. Is that close enough to 0.5 to be acceptable, or should we claim that a difference of 0.013 refutes our propensity hypothesis?

We can also present these difficulties by means of some simple probability calculations. Suppose our hypothesis h is as in the previous paragraph, and that the underlying conditions S are repeated n times with the event A occurring r times. Assuming that the repetitions are independent, we have

$$\text{Prob}(r/n) = {}^{n}C_{r}p^{r}(1-p)^{n-r} \tag{19.1}$$

So, however long our sequence of repetitions (however big n is) and however many times the event A occurs (whatever the value of r), the observed frequency of A (r/n) will always have a positive probability. It will not be strictly ruled out by our hypothesis h. In other words, h cannot be falsified by observed data, and so, according to Popper's well-known criterion, would seem to be a metaphysical rather than a scientific hypothesis.

The difficulty about falsifying probability statements was noted by Popper early on, and he gives the following statement of it in *The Logic of Scientific Discovery* (1934/1972: p. 146):

an account of Miller's attempt in his 1994 to retain both the causal directedness of propensities and the standard axioms of probability. It is also valuable to consider the relations between propensity and causality in the context of Bayesian networks. An attempt in this direction is to be found in my 2002.

> The relations between probability and experience are also still in need of clarification. In investigating this problem we shall discover what will at first seem an almost insuperable objection to my methodological views. For although probability statements play such a vitally important rôle in empirical science, they turn out to be in principle *impervious to strict falsification*. Yet this very stumbling block will become a touchstone upon which to test my theory, in order to find out what it is worth.

Popper's answer to this difficulty consists in an appeal to the notion of methodological falsifiability. Although, strictly speaking, probability statements are not falsifiable, they can nonetheless be used as falsifiable statements, and in fact they are so used by scientists. Popper puts the matter thus (1934/1972: p. 191):

> ... a physicist is usually quite well able to decide whether he may for the time being accept some particular probability hypothesis as "empirically confirmed", or whether he ought to reject it as "practically falsified" ...

Popper's approach has been strongly vindicated by standard statistical practice. Working statisticians are constantly applying one or other or a battery of statistical tests, such as the chi-square test, the t-test or the F-test. The procedure in any such statistical test is to specify what is called a 'rejection region', and then regard the probabilistic hypothesis under test (H say) as refuted if the observed value of the test statistic lies in this rejection region. Now there is always a positive probability (called the 'significance level' and usually set around 5%) of the observed value of the test statistic lying in the rejection region when H is true. Thus H is regarded as refuted, when, according to strict logic, it has not been refuted. This is as much as to say that H is used as a falsifiable statement, even though it is not, strictly speaking, falsifiable, or, to put the same point in different words, that methodological falsifiablility is being adopted.[8]

Let us now apply the methodological falsifiability implicitly used by statisticians to our example of the hypothesis h that the propensity of A = p. From h, we deduced equation (19.1) above. This is a binomial distribution which for large n tends to the normal distribution. If we base a statistical test on this normal distribution using the standard level of significance of 5%, we can show that there is a 95% probability of the relative frequency of A, that is r/n, lying in the interval

$$[p - 1.96(p(1-p)/n)^{1/2}, p + 1.96(p(1-p)/n)^{1/2}] \qquad (19.2)$$

Applying the statisticians' methodological falsifiability, we regard it as practically certain that the r/n will lie in the interval (2), and so take the hypothesis h to be refuted if the observed value of r/n lies outside this interval. In this way a connection is established between theoretical probabilities and observed frequencies without probability ever being defined in terms of limiting frequency as it was by von Mises and other frequency theorists.

We can illustrate this procedure by a simple example. Suppose that p equals ½. Then interval (2) becomes

$$[1/2 - 0.98/n^{1/2}, 1/2 + 0.98/n^{1/2}] \qquad (19.3)$$

[8] I have tried to formulate methodological falsifiability in terms of what I call a 'falsifying rule for probability statements', and to compare this rule to the procedures of statistical testing used in practice. An account of these matter is to be found in my 1971, or in my 1973: Part III pp. 161–226.

Table 19.1

Author	Number of tosses	Allowable Deviation	Relative Frequency of Heads	Difference between r.f. and 0.5
Gillies	2000	±0.022	0.487	-0.013
Buffon	4000	±0.015	0.507	+0.007
K. Pearson	12000	±0.009	0.502	+0.002
K. Pearson	24000	±0.006	0.501	+0.001

Earlier I considered the example of h with p = ½ and 2,000 repetitions, and asked the question of how close to 0.5 the observed frequency should be in order to be compatible with h. If we adopt the significance level of 5%, as is customary in statistical practice, we can now calculate the answer from (3) by putting n = 2000. This calculation shows that the allowable deviation from 0.5 at this level of significance is ±0.022. So a result for r/n of 0.487, that is a deviation of 0.013, is in fact allowable and does not refute h at the 5% level.

This example is based on a coin-tossing experiment which I performed some years ago with an ordinary old penny. Table 19.1 gives the results of this, and other similar coin tossing experiments carried out by Buffon and Karl Pearson.

In each case the allowable deviation is calculated at the 5% level of significance, and in each case the observed frequency is compatible with the hypothesis that the coin was a fair one. These figures show in a vivid way how theoretical propensities can be related to observed frequencies by methodological falsifiability without the need for any definition of probability in terms of limiting frequency.

Let me now say something more about the assumption that propensities satisfy the Kolmogorov axioms. Assuming methodological falsifiability, we can test empirically specific propensity hypotheses, such as the claim that P(A|S) = p, either refuting or confirming the hypothesis. However, the assumption that propensities satisfy the Kolmogorov axioms is at too high a theoretical level to be tested in this way, even assuming methodological falsifiability. What can be said, however, is that the assumption that propensities satisfy the Kolmogorov axioms, together with methodological falsifiability, has led to the development, in a great variety of different areas, of propensity hypotheses which are empirically testable, and which have been tested and confirmed. This provides an indirect empirical justification for the assumption that propensities satisfy the Kolmogorov axioms, and for the usefulness of the methodological falsifiability adopted by statisticians in their practice.

Hacking in his 1965 book *Logic of Statistical Inference* develops what could also be called a long-run propensity theory, but it differs in one interesting respect from the long-run propensity theory which we have sketched in this section. Hacking uses the term 'chance' rather than 'propensity'. However, he says (1965: p. vii):

> Chance is a dispositional property: it relates to long run frequency much as the frangibility of a wine glass relates to whether the glass does, or did, or will break when dropped.

Hacking also points out (1965: pp. 10–11) that his term 'chance' has a similar meaning to Popper's 'propensity'.

Hacking rejects von Mises' definition in terms of limiting frequency, and suggests instead that chance should be given a postulational definition. Part of this postulational definition

is that chances satisfy the Kolmogorov axioms. However, something more is needed. As Hacking says (1965: p. vii):

> Although the Kolmogoroff axioms help to define chance they are not enough. … we shall need further postulates to define chances and to provide foundations for inferences about them.

So far, Hacking's theory of chance is very similar to the long-run propensity theory expounded in this section, but now comes the difference. The further postulate added by the long-run propensity theory of this section is that of methodological falsifiability. Hacking, however, suggests adding postulates which specify when statistical hypotheses are supported by data. In his 1965 book, he develops a logic of empirical support and shows how it can be related to standard views of statistical inference. In Hacking's version of a long-run propensity theory, therefore, the emphasis is on empirical support or confirmation rather than on empirical refutation or falsifiability.

This is another illustration of the variety of different theories which come under the heading of 'propensity'. I hope in this chapter to have given the reader some idea of this variety, and of the complex group of related problems which have not been fully resolved and which will undoubtedly provide a stimulus for further developments.

Acknowledgments

I would like to give particular thanks to Alan Hájek with whom I had very lengthy discussions about the propensity theory. These started over dinner on a visit he paid to London, and then continued by email. The points he raised have led me to revise my views on the propensity interpretation so that the present chapter differs in several respects from my previous writings on this subject.

References

Ayer, A. J. (1963) Two Notes on Probability. In *The Concept of a Person and Other Essays*. pp. 188–208. London: Macmillan.

Fetzer, J. H. (1981) *Scientific Knowledge: Causation, Explanation, and Corroboration*. Boston Studies in the Philosophy of Science. Dordrecht: Reidel.

Fetzer, J. H. (1982) Probabilistic Explanations. *Proceedings of the Biennial Meeting of the Philosophy of Science Association.* Vol. 2. pp. 194–207. Chicago, IL: University of Chicago Press.

Fetzer, J. H. (1988) Probabilistic Metaphysics. In Fetzer, J. H. (ed.) *Probability and Causality*. pp. 109–32. Dordrecht: Reidel.

Fetzer, J. H. (1991) Critical Notice: Philip Kitcher and Wesley C. Salmon (eds.) *Scientific Explanation*; and Wesley C. Salmon, *Four Decades of Scientific Explanation*. *Philosophy of Science*. 58. pp. 288–306.

Fetzer, J. H. (1993) Peirce and Propensities. In Moore, E. C. (ed.) *Charles S. Peirce and the Philosophy of Science*. pp. 60–71. Tuscaloosa and London: University of Alabama Press.

Giere, R. N. (1973) Objective Single Case Probabilities and the Foundations of Statistics. In Suppes, P. , Henkin, L., Joja, A., and Moisil, Gr. C. (eds.) *Logic, Methodology and Philosophy of Science*. Vol. IV. pp. 467–83. Amsterdam: North-Holland.

Gillies, D. A. (1971) A Falsifying Rule for Probability Statements. *British Journal for the Philosophy of Science*. 22. pp. 231–61.

Gillies, D. A. (1973) *An Objective Theory of Probability*. London: Methuen.

Gillies, D. A. (2000a) Varieties of Propensity. *British Journal for the Philosophy of Science*. 51. pp. 807–35.

Gillies, D. A. (2000b) *Philosophical Theories of Probability*. London and New York: Routledge.

Gillies, D. A. (2002) Causality, Propensity, and Bayesian Networks. *Synthese*. 132. pp. 63–88.

Hacking, I. (1965) *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hájek, A. (2002/2011) Interpretations of Probability. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/probability-interpret/ [Accessed 11 Oct 2015.]

Howson, C. and Urbach, P. (1989) *Scientific Reasoning: The Bayesian Approach*. Lasalle, IL: Open Court.

Humphreys, P. (1985) Why Propensities Cannot Be Probabilities. *Philosophical Review*. 94. pp. 557–70.

Kant, I. (1787/1958) *Critique of Pure Reason*. Translated from the German by N. K. Smith. 2nd edition. London: Macmillan.

Kolmogorov, A. N. (1933/1956) *Foundations of the Theory of Probability*. 2nd English edition. New York: Chelsea.

Mellor, D. H. (1971) *The Matter of Chance*. Cambridge: Cambridge University Press.

Miller, D. W. (1994) *Critical Rationalism: A Restatement and Defence*. Chicago and Lasalle, IL: Open Court.

Miller, D. W. (1996) Propensities and Indeterminism. In O'Hear, A. (ed.) *Karl Popper: Philosophy and Problems*. pp. 121–47. Cambridge: Cambridge University Press.

Peirce, C. S. (1910/1957) Notes on the Doctrine of Chances. In *Essays in the Philosophy of Science*, *The American Heritage Series*. pp. 74–84. Indianapolis, IN and New York, NY: Bobbs-Merrill.

Popper, K. R. (1934/1972) *The Logic of Scientific Discovery*. 6th revised impression. London: Hutchinson. (1^{st} impression 1959. First published in German.)

Popper, K. R. (1957) The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory. In Körner, S. (ed.) *Observation and Interpretation*, *Proceedings of the Ninth Symposium of the Colston Research Society*. pp. 65–70, 88–9. University of Bristol.

Popper, K. R. (1959) The Propensity Interpretation of Probability. *British Journal for the Philosophy of Science*. 10. pp. 25–42.

Popper, K. R. (1967) Quantum Mechanics without 'the Observer'. In Bunge, M. (ed.) *Quantum Theory and Reality*. pp. 7–44. New York: Springer.

Popper, K. R. (1983) *Realism and the Aim of Science*. London: Hutchinson.

Popper, K. R. (1990) *A World of Propensities*. Bristol: Thoemmes.

Runde, J. (1996) On Popper, Probabilities and Propensities. *Review of Social Economy*. LIV 4. pp. 465–85.

Salmon, W.C. (1979) Propensities: A Discussion Review. *Erkenntnis*. 14. pp. 183–216.

von Mises, R. (1928/1961) *Probability, Statistics and Truth*. 2nd revised English edition. London: George Allen & Unwin.

CHAPTER 20

BEST SYSTEM APPROACHES TO CHANCE

WOLFGANG SCHWARZ

20.1 Introduction

In the early 20th century, developments in quantum physics suggested that the present state of the world determines only certain probabilities for future states. This type of probability is often called *chance*. Unlike degree of belief, chance does not vary from person to person. Unlike degree of evidential support, it is contingent and not relative to a body of evidence. There is an intriguing normative connection between chance and degree of belief (discussed in Section 20.5 below), but this connection does not tell us what chance is, for we can hardly assume that the basic laws of physics involve irreducibly normative and psychological notions. So what is chance?

The question is philosophically controversial in part because chance is a modal phenomenon. The fact that an event has chance x appears to entail neither that it takes place nor that it doesn't take place; chance therefore seem to point beyond the history of actual outcomes and events to a sphere of mere possibilities. Indeed, chance looks like a graded form of physical necessity: the higher the chance of an event, the closer it is to physical necessity.

Philosophers are deeply divided on the status of modal truths in general and nomic truths in particular – truths related to physical possibility and necessity. Besides chance, prominent nomic phenomena include causation, counterfactuals, dispositions and laws of nature. One approach to the nomic, going back to Hume and further to the medieval nominalists, holds that nomic facts about what *could* or *would* or *must* are always reducible to facts about what *is*. For the most part, Humeans do not deny the reality of nomic phenomena; they agree that there are laws of nature, chances, dispositions, etc. But they maintain that these things are derivative, determined by more fundamental, non-modal elements of reality. Thus Humeans might identify laws of nature with regularities in the history of physical events, and chances with relative frequencies. This ensures that whenever two possible worlds agree in non-nomic respects, they also agree with respect to laws and chance.

On the opposite side are those who follow Aristotle and maintain that nomic elements are woven right into the fabric of reality, as Aristotle's "substantial forms" perhaps, or as primitive powers or primitive laws. With respect to chance, a characteristic anti-Humean view is the primitive propensity account, defended e.g. in Mellor 1971 and Giere 1973 (see also Donald Gillies' chapter (19) in this volume). Here chance is assumed to be a basic physical quantity not unlike mass or charge. Just as an atom's mass grounds a disposition to interact in certain ways with its environment, an atom's propensity to decay within a certain interval of time grounds a "partial disposition" to decay within that time.

Humeans and anti-Humeans are generally easy to identify, although it is hard to state their disagreement precisely, without relying on controversial physical and metaphysical assumptions. Humeanism is sometimes characterized (e.g. in Lewis 1986b, pp. ix–xi, Lewis 1994, 225f., and Loewer 2012, p. 116) as the view that all truths supervene on the distribution of fundamental categorical properties over points or regions of spacetime (or some other physically basic space), where a categorical property is one whose instantiation in a region entails nothing about the instantiation of fundamental properties outside that region. However, this would not exclude primitive propensities, since instantiation of a propensity does not strictly entail anything about other times or places.

The present chapter is concerned not with Humeanism in general, but with a particular Humean analysis of chance: the "best system account", first proposed by David Lewis in Lewis 1986b, p. 128 and Lewis 1994. The analysis can be seen as a sophisticated descendant of the frequency interpretation (see La Caze's chapter (16) in this volume). Unlike simple frequentism, the best system account allows chances to come apart from actual frequencies: if there is a 50 percent chance that radium atoms decay within 1600 years, it does not follow that exactly 50 percent of the atoms really decay within that time (and therefore that there is an even number of radium atoms). Rather, chances are identified with probabilities in ideal physical theories whose aim is to provide a kind of summary statistic of actual outcomes; getting close to the frequencies is one virtue of probabilistic theories, but it trades off against other virtues such as comprehensiveness and simplicity.

20.2 Laws and Chance

Lewis's best system account of chance begins as a best system account of laws, also known as the Mill-Ramsey-Lewis account, owing to its origins in Mill 1843, Ramsey 1978, and Lewis 1973. If we look at popular theories in fundamental physics – the theory of general relativity, say, or the standard model of particle physics – we find that they predict a staggering variety of facts with very high precision, based on relatively few and simple assumptions. That is no coincidence. What physicists aim for is precisely to find simple principles that allow a unified explanation for all sorts of complex phenomena. Of course our present theories may turn out to be false; some physicists also hope to find an even more unified and comprehensive "theory of everything". Suppose there is such a theory – whether or not we will ever find it. Suppose it makes no false predictions, and there is no other theory that accounts for an even greater range of facts with even simpler rules. On the best system account, it follows that the rules of this best theory are the true laws of nature. Lewis (1994, p. 231f.) succinctly states the general analysis:

> Take all deductive systems whose theorems are true. Some are simpler, better systematized than others. Some are stronger, more informative, than others. These virtues compete: an uninformative system can be very simple, an unsystematized compendium of miscellaneous information can be very informative. The best system is the one that strikes as good a balance as truth will allow between simplicity and strength.…A regularity is a law iff it is a theorem of the best system.

In worlds like ours, there are especially salient regularities linking earlier and later states of physical systems. Perhaps these regularities can be captured by simple equations expressing a functional relationship between the state of isolated systems at one time and their past and future evolution. But perhaps things are not so simple. Perhaps a physical system in state X evolves sometimes into state Y and sometimes into Z, without any intrinsic difference in the initial states or their histories. In this case, it might nevertheless be useful to learn that X states are followed by either Y states or Z states. Moreover, it might be valuable to know something about the proportions: is one of the outcomes much more common than the other? A good theory might then specify probabilities for a system in state X to turn into Y or Z. In general, if the history of a world reveals a pervasive but "noisy" dependence between two quantities, then a good theory might specify a probabilistic connection between these quantities to convey information about their distribution. Lewis suggests that these probabilities are the chances: "the chances are what the probabilistic laws of the best system say they are" (1994, p. 234).

This may look circular, but it is not. The idea is that one can evaluate probabilistic theories without first interpreting their probabilistic statements:

> As before, some systems will be simpler than others. Almost as before, some will be stronger than others: some will say either what will happen or what the chances will be when situations of a certain kind arise, whereas others will fall silent both about the outcomes and about the chances. And further, some will fit the actual course of history better than others. That is, the chance of that course of history will be higher according to some systems than according to others.…The virtues of simplicity, strength, and fit trade off. The best system is the system that gets the best balance of all three. (1994, p. 234)

To illustrate, imagine a world consisting of nothing but a short sequence of binary events ("coin tosses"), with outcomes 0111000011. Let T_1 be a theory according to which the events are independent with both 1 and 0 having constant probability $1/2$. Theory T_2 assigns probability $9/10$ to 1 and $1/10$ to 0. The total sequence then has probability $(1/2)^{10} \approx 0.001$ according to T_1 and probability $(9/10)^5(1/10)^5 \approx 0.000005$ according to T_2. In general, the closer a theory's probability assignments match the relative frequencies, the greater the probability of the actual sequence and therefore the better the theory's fit. Another theory, T_3, might relax the independence assumption and say that the probability of 1 is $2/3$ immediately after an occurrence of 1 and otherwise $1/3$. The sequence then has probability around 0.002; but this increase in fit comes at a cost in simplicity.

The real value of probabilistic theories becomes apparent only in more complex worlds. Imagine we knew the precise decay times for every atom in the history (past and future) of the world. How could we summarize this information? Listing the actual frequencies for every type of atom and every interval of time would be as unwieldy as listing the individual decay times. One could instead give the average decay time for each type of atom, or the average and the standard deviation. Much better, one could adopt the language of

probability and specify an exponential distribution of decay times for each element, with the understanding that the actual frequencies approximately fit that curve. This would not allow reconstructing the exact actual pattern of decay times, because many alternative patterns would be summarized by the very same distribution, but it would convey a lot of information about the pattern in a very brief and elegant manner.

To get a full analysis of chance, we would now have to spell out the ordering that determines which theory counts as best relative to a possible history of events. It is not crucial to the best system approach how exactly these details are filled in. Nevertheless, it is worth having a closer look at some issues this involves.

20.3 Comparing Theories

The best system account defines the chances at a world as the probabilities that figure in the best theory for that world. To this end, Lewis identifies theories with arbitrary deductively closed sets of sentences ("systems") in some fixed language *L*. To avoid the choice of a language, it is tempting to instead identify theories with sets of models. But this might cause difficulties with probabilistic theories: what should count as a model of a probabilistic theory, given that the probabilities are as yet uninterpreted? I will therefore stick with Lewis's somewhat anachronistic syntactical conception of theories.

Lewis mentions three criteria of goodness for such theories: simplicity, strength, and fit. Let's take these in turn, beginning with simplicity. If theories are sets of sentences, simplicity is naturally understood as syntactical complexity. Roughly speaking, the fewer axioms are needed to generate a theory, and the simpler their logical and mathematical form, the simpler the theory.

Evidently, any such measure is highly sensitive to the chosen language. If *T* is a complicated (non-probabilistic) theory axiomatized by 147 long and unrelated principles, we can define an atomic predicate *F* to be true of an object *x* iff *x* exists in a world where those 147 principles are true; *T* can then be translated into the very simple theory "everything is *F*" (see Lewis 1983, p. 42). Hence Lewis requires that before entering the competition, all theories must be translated into a common language *L* in which all basic non-logical terms stand for fundamental properties (or magnitudes) such as mass or charge. Gerrymandered predicates such as *F* are forbidden.

This move has raised some eyebrows (see e.g. van Fraassen 1989, pp. 52–5, Loewer 1997, sec.3, Cohen and Callender 2009). For how do we know which properties are fundamental? If two theories postulate different sets of fundamental properties, how can we tell which is right? Lewis holds that fundamental properties are also related to similarity and duplication: two objects are perfect duplicates iff they agree in the distribution of fundamental properties and relations over their parts (see Lewis 1986a, p. 61f.). But arguably this does not rule out the skeptical possibility that scientists might settle on apparently simple regularities that are not the true laws because they are not very simple any more when translated into the language of objectively fundamental properties.

Uncomfortable with this consequence of Lewis's proposal, several advocates of the best system account have suggested an alternative on which every theory may provide its own

inventory of fundamental properties (see Loewer 2007b, Cohen and Callender 2009). On this view, there is no objective, theory-independent standard for comparing systems with different inventories; hence there is nothing objectively wrong with a system that encodes many (or even all) truths in a single statement "everything is *F*". Perhaps the problem with this system is only that it does not strike us as very perspicuous.

What shows up in this discussion is the other great divide in the metaphysics of science, between scientific realism and anti-realism. Lewis assumes that the world has an objective, mind-independent structure, given by the distribution of fundamental properties and relations. Scientific theories try to uncover patterns in this structure, but there is always a skeptical possibility that they fail, that the lines drawn by science do not carve nature at its joints. Cohen and Callender, on the other hand, reject the idea of objectively natural or fundamental properties: objectively, the negatively charged particles have no more in common than an electron, a goat, and the Atlas mountains. Our preference for theories expressed in terms of charge, mass, etc. therefore only reflects idiosyncratic features of our interests and upbringing. The best system account can be made to fit both the realist and the anti-realist picture.

Lewis's second criterion is strength. An obvious way to understand strength is in terms of excluded possibilities; the more possibilities a system excludes, the greater its strength. It is not clear what exactly this means given that the excluded possibilities are typically infinite. There are also problems with probabilistic theories, which don't seem to exclude anything – especially if the probabilities are uninterpreted. On reflection, the whole idea is on the wrong track anyway. Consider two worlds w_1 and w_2; w_1 contains a lot of *F*s, all of which are *G*; w_2 is like w_1 except that it contains very few *F*s (perhaps none). In this case, "all *F*s are *G*" is a better candidate to be a law of w_1 than a law of w_2, since it provides much more valuable information about w_1. It follows that strength is not an intrinsic aspect of theories. Whether one theory is stronger than another depends on which world is under consideration. Roughly speaking, the more *F*s there are in a world, and the fewer *G*s, the stronger is a law "all *F*s are *G*s" (or "the probability of an *F* being *G* is *x*") with respect to that world.

John Earman has often pointed out (e.g. in Earman 1984, Earman 1986, and Earman and Roberts 2005) that the strength scientists value in physical theories is actually more specific. Modern physical theories typically provide differential equations which, combined with suitable boundary conditions concerning (say) a system's state at a particular time, allow computing the system's state at earlier and later times. What we value is strength in these dynamical laws. The ideal limit are deterministic laws, which allow complete reconstruction of a system's history from information about a single time. By contrast, strength with respect to boundary conditions is much less of a virtue, even if it would make a system more informative. For an extreme example, imagine a world with just a few thousand particles and deterministic dynamical laws. For each time, there is a statement that describes the exact state of all particles at that time. Adding the simplest among those statements to the dynamical laws yields a maximally informative system. Nevertheless, we are not inclined to deem this a very good physical theory, nor would we say that in such a world, all facts are physically necessary.

Lewis explicitly mentions that a theory may involve statements of particular fact, in which case only the "regularities" it entails count as laws (see Lewis 1986b, p. 123). He does not explain what makes something a regularity (obviously this is not a matter of syntax), but we

can assume that boundary conditions are excluded. To some extent, it is then unimportant for the analysis of laws and chance if systems may include boundary conditions or if instead their strength is measured by their informativeness when combined with external boundary conditions. Either way, the boundary conditions do not come out as laws or as physically necessary.

One may still wonder why a good system should include dynamical laws at all, or dynamical laws plus boundary conditions. If all that counts is informativeness or logical strength that is hard to explain. But we've seen that informativeness by itself is not the right standard to begin with. The right standard is world-relative, and – well – it privileges strength in dynamical laws. On the best system account, the standards *define* the notions of laws and chance. To the question "what makes those standards right?" there is no substantive, metaphysical answer. We could have used a different notion of (quasi-)laws defined by other standards. That in fact we are more interested in systems with strong dynamical laws is easy to understand: we often have information about the present (or recent past) of physical systems, and want to know more about their future and distant past. A system of laws that applies (approximately) to (more or less) isolated subsystems of the universe, in all sorts of boundary conditions, is a very useful thing to know.

Lewis's third criterion for theories is "fit". Lewis suggests that this is measured by the probability a theory assigns to the entire history of the world. This is a natural idea, but it suffers from a technical difficulty, known as the *zero-fit problem*: good theories often assign zero probability to the entire history of a world, either because the world contains infinitely many chance events, or because the space of outcomes for individual events is infinite. For example, if the decay probabilities for radium atoms follow an exponential distribution, the probability of any particular atom decaying at any particular time is zero. Lewis (1980) did regard this as a merely technical problem, pointing at Bernstein and Wattenberg (1969) for a solution based on probability theories that allow for infinitesimal values. Elga (2004) argues that these infinitesimals cannot be relied upon to yield the desired ordering on theories in terms of fit. Drawing on Gaifman and Snir (1982), he instead suggests defining fit by looking at the probabilities a theory assigns to a carefully chosen set of truths about a history, rather than the conjunction of all truths.

Another problem with Lewis's (and to some extent Elga's) proposal is that a good theory may assign no probability at all to the total history of the world. A stochastic differential equation, for example, directly specifies probabilities only to transitions between states. If the world has an initial state, one could perhaps compute a probability for the entire history relative to this starting point; but this is hardly a general solution.

At this point, it might be a good idea to look at "goodness of fit" measures developed in statistics, such as the chi-squared test of classical statistics or one of its Bayesian counterparts. Interestingly, these measures usually include simplicity constraints to avoid overfitting. This illustrates that Lewis's three criteria need not be treated as independent: the ordering that defines the best system doesn't need to come about by somehow balancing three independent scores. In the context of the best system approach, an especially noteworthy approach to statistical model selection is the framework of "Minimum Description Length", introduced in Rissanen (1978). Lewis's Humean perspective on the relationship between theories and the world bears a striking resemblance to Rissanen's perspective on the relationship between models and data. Rissanen rejects the idea that the goal of statistical inference is to find the true probability distribution that "generated" the data (see especially

Rissanen 1989); likewise, Lewis rejects the idea that the pattern of events in the world is "generated" by some hidden, underlying probability distribution; objective probabilities simply state noisy regularities in the pattern of events.

Scientific practice also suggests further conditions on good theories. For example, physicists generally prefer theories with fewer primitive constants, even if reducing the constants comes at a slight cost in simplicity. To measure this, it can be useful to consider not only an individual theory, but the parametric family of theories that results by leaving the value of constants open and checking how many patterns of outcomes can be made to fit by tweaking the constants: the more, the worse (see Myung et al. 2000).

However, we should not assume that the standards for good theories relevant to the best system account precisely match the standards used in science (although Lewis says so in Lewis 1986b, p. 123). Model selection criteria are often constrained by computational complexity. Scientists also tend to prefer theories that are conservative with respect to earlier theories, presumably on the ground that it is much easier to see things correctly if one is standing on the shoulders of giants. These features seem out of place in the analysis of laws and chance. When scientists invoke criteria of statistical model selection, they consider the adequacy of theories in the light of a limited set of observations, with an eye to predicting the unobserved. By contrast, for the best system account we compare a theory's probabilities with the total history of the world (observed and unobserved), in order to define a goodness order on the space of theories which then defines the notions of laws and chance. The only direct constraint here is the plausibility of the resulting analysis. To which we turn now.

20.4 Playing the Chance Role

Best system accounts identify chance with certain features of the total history of events in a world. It is relatively uncontroversial that these features exist. But can they play the role of chance in scientific and philosophical thought?

The question, it turns out, is ill posed. There is no such thing as *the* role of chance in scientific and philosophical thought. A symptom of this is the debate over whether chance is compatible with determinism: if the laws of nature are deterministic, can there still be non-trivial chances pertaining for instance to the outcomes of coin tosses? Some, including Popper (1982, p. 105), Lewis (1986b, p. 120), Hájek (1997, p. 221), and Schaffer (2007), take the answer to be obviously "no". Others, including Levi (1990), Loewer (2001), and Hoefer (2007) deem it equally non-negotiable that the answer is "yes". Philosophers also disagree over whether chance is essentially dynamical, linking earlier to later states, and whether past events can still be chancy. One can hardly expect any one analysis of chance to satisfy all these contradictory constraints.

Some aspects of chance, at least, are uncontroversial. First of all, most authors agree that chance (in the sense of physical probability) satisfies some form of the probability axioms. On the best system account, this could simply be the consequence of a stipulation that good systems should be not only logically consistent, but also probabilistically coherent. More ambitiously, one might argue that it follows from the fact that for every probabilistically incoherent theory there is a coherent theory that is guaranteed to have greater fit no matter

what the world is like (see Joyce (2009) for observations of this type, although not applied to chance).

Secondly, it is usually agreed that chances figure in laws of nature and are thus linked to other nomic phenomena such as causation and counterfactuals. By itself, the best system account of chance does not establish this; however, the account naturally goes with a corresponding account of laws and other nomic phenomena, which secures the desired links.

Scientific practice requires a close connection between chance on the one hand and symmetries and frequencies on the other. In particular, relative frequencies in long series of trials are assumed to be close to the chances. This substantive fact is sometimes mistaken as a consequence of the Laws of Large Numbers; but the Laws hold for all probability functions whatsoever, even ones that are completely out of tune with the frequencies. On the best system account, the alignment between chance and frequencies is not too surprising, since a good system must have high fit. Similarly, invariance under symmetries is a plausible virtue of physical theories, so it is unsurprising that the chances should respect it.

This third feature of chance leads to a fourth: chances can be discovered by the methods of science. Again, the best system account makes this understandable, since one can reasonably hope that the methods of science get us close to the ideal theories whose probabilities are identified with the chances. Of course, there is no logical guarantee of success. Demanding as much would be implausible anyway. It is logically possible that our observations are systematically skewed, that relative frequencies in observed samples are radically different from the overall frequencies, and so on. If these skeptical possibilities show that science cannot discover the chances (on the best system account), they show that science cannot discover general truths at all.

Two further important aspects of chance concern the distribution of chance events in the world. First (or rather, fifth), the chance of an outcome is typically constant across intrinsically similar situations: if two systems perfectly agree in the distribution of mass, charge, spin, etc., we expect them to also agree about the dynamical chances (see Arntzenius and Hall 2003). On the best system account, this makes sense: if the chances varied widely and unsystematically between intrinsically identical situations, they could hardly be specified in a simple and elegant theory. On the other hand (sixth), we expect disorder in the *outcomes* of chance: chance events usually have a random-looking distribution. This is also predicted by the best system account: if the outcomes of a process come in a conspicuous pattern (such as 1010101010…), a good theory would not need to resort to probabilities.

All these issues await more detailed investigation, spelling out under what conditions and on what background assumptions the best system account can vindicate the stated facts about chance. But it is fair to say that the prospects look considerably better than e.g. on the primitive propensity account. If chance is a primitive quantity logically independent of the distribution of other quantities, then it is rather mysterious – to put it mildly – why this quantity should be linked to frequencies and symmetries among other quantities, why it doesn't vary when those other quantities are held constant, or why chance events come in disorderly patterns.

A further advantage of the best system approach is that it can vindicate many of the more controversial views about chance – even when these appear to contradict one another. This is because it allows for different species of chance. Let's reserve the title *fundamental dynamical chance* for probability functions in a Lewisian best system that pertain to later

states of a world given an earlier state. Fundamental dynamical chance exists only in indeterministic worlds. Since it is always conditional on a given (earlier) state of the world, it appears to evolve over time by conditioning on history, and can be stated in the form of "history-to-chance conditionals" – just as for example Lewis (1980) and Schaffer (2007) intuit. But the best system approach allows also for other ways in which probabilities could find themselves in a best system. For example, a best system might involve a non-dynamical probability distribution over initial states of the universe, as in some versions of statistical mechanics or Bohmian mechanics. This type of chance is compatible with deterministic dynamical laws. Whether or not it deserves the honorific label "chance", it is an objective physical quantity, just as Loewer (2001), Hoefer (2007), and others insist.

The best system account can even vindicate many claims of the propensity interpretation. As we saw, in well-behaved worlds with fundamental dynamical chances (in the sense of the best system account), the chances are plausibly *intrinsic to experimental setups* insofar as duplicating an experiment also duplicates the chances. Since the chances are given by "partial laws" which, instead of saying that all *F*s are *G* merely assign a probability to *G* given *F*, one might say that the condition *F partially necessitates* manifestation of *G*, so that the relevant system has a *partial disposition* to produce *G*. Like propensities, fundamental dynamical chances are *forward-directed* and *essentially conditional*. The only point where the best system account must depart from propensity accounts is when propensities are declared metaphysically fundamental.

Three more aspects of the chance role deserve special attention. One is the link between chance and rational belief; another is the apparent independence of chances from actual outcomes. These will be discussed in Sections 20.5 and 20.6, respectively. The third is the objectivity of chance.

This is often regarded as a problem for best system accounts. Chances (and laws of nature) are supposed to be objective, mind-independent features of the world. But don't the standards for good theories, in particular the standards of simplicity, depend on us? Humeans with anti-realist or pragmatist sympathies might endorse this consequence and reject the idea of mind-independent laws and chances as targets of scientific inquiry. As a philosopher of a more realist bent, Lewis (1994, pp. 232f.) took the objection more seriously. He offered two lines of response. First, he argued that it is not just a matter of taste which theories we deem more or less simple (when translated into the language of fundamental properties). Indeed, we think simpler theories are more likely to be true; this would make no sense on the assumption that the relevant standards of simplicity merely reflect epistemically irrelevant facts about human psychology. In addition, Lewis suggested that our concepts of laws and chance only have clear application in worlds where a unique best system is significantly ahead of the competition, so that the laws and chances do not depend on fine details of the comparison.

But this does not fully resolve the worry. Even if there are objective standards for what to believe about the world given limited observations, it does not follow that there are objectively privileged standards for how best to summarize the entire history of the world. Consider again our preference for strength in dynamical laws. This arguably reflects our epistemic limitations and interests. Does this not undermine the objectivity of laws and chance? The worry is easy to sense, but harder to make precise. After all, the best system account does not define chance indexically as the probability in whatever theory comes out best according to our standards, whatever they might be. Rather, a fixed set of standards is

used in the definition. To be sure, these standards come from us – but where else should they come from? The same is true for all terms. Whether something falls in the extension of "positively charged" (say) also depends on contingent facts about how we apply those words; it does not follow that positive charge is not an objective, mind-independent matter. However, there is a difference. On the realist picture, our notion of positive charge picks out a metaphysically privileged, fundamental joint in nature. Not so our notion of chance, if the best system account is right: it picks out something objective, but nothing *objectively special*. Chance is special only to creatures like us.

20.5 Chance and Credence

Perhaps the most striking feature of chance is its connection to rational belief. As a first stab, the connection might be expressed as follows. If Cr is a rational credence function and $Ch_t(A) = x$ a proposition specifying that the chance of A at time t is x, then

$$Cr(A/Ch_t(A) = x) = x,$$

provided $Cr(Ch_t(A) = x) > 0$. For example, knowing that a given coin has chance 1/2 of landing heads, one should have equal confidence in heads and tails. Many variations of this principle have been put forward, under various names. A problem with the present formulation is that it doesn't take into account the possibility that the agent has further information about the relevant proposition A, as when she has already seen the coin land heads (and t is a time in the past). On the other hand, if the coin toss lies in the future, it is hard to imagine any evidence that would affect one's rational confidence in heads and tails, given knowledge that the coin is fair. Taking these two observations into account leads to a version of the chance-credence link which Lewis (1980) dubbed the *Principal Principle*. Again, there are slightly different ways of stating the principle; Lewis himself gave several non-equivalent formulations. The first and best known goes in essence as follows. If Cr is any rational *initial* credence function, $Ch_t(A) = x$ a proposition saying that the chance of A at t is x, and E arbitrary *admissible* information, then

$$Cr(A/Ch_t(A) = x \wedge E) = x,$$

provided $Cr(Ch_t(A) = x \wedge E) > 0$. "Admissible" information is the kind of information one can reasonably acquire ahead of the relevant time t.

The Principal Principle, in some form or another, is crucial to our understanding of chance. It also serves as a powerful touchstone for candidate interpretations: chance can be identified with an objective quantity X only if it is plausible that X guides rational credence in the way demanded by the Principle. This is where Lewis, until the early 1990s, saw a fatal problem with his best system account – for he had discovered that the analysis is incompatible with the Principal Principle (see Lewis 1980, pp. 111f. and Lewis 1986b, pp. 129–31). He famously called this the "Big Bad Bug" for his whole Humean metaphysics (see Lewis 1986b, p. xiv).

The problem can be illustrated in a simple coin-toss universe. Imagine a world consisting of 10 binary events, and consider again the theory T_1 that takes the outcomes to be

independent with constant probability 1/2. The sequence 1111111111 then has probability $(1/2)^{10}$, just like 0111000011. But 1111111111 is not best systematized by T_1. So T_1 assigns positive probability to situations that are incompatible with T_1 being the best system. In general, no theory on which the chance of all-1 is less than 1 best systematizes the all-1 history. Thus for any x strictly between 0 and 1, $Cr(\text{all-1}/Ch(\text{all-1}) = x) = 0 \neq x$.

The phenomenon at issue is best known from so-called "undermining futures". Let Z be a proposition according to which most radium atoms from now on decay within a few years. Suppose this proposition, although false, has positive chance x. In worlds where Z is true, the history of actual events may well be different enough from the actual history to give rise to different chances (on the best system account). In particular, the chance of Z in such worlds may not be x. On the supposition that the chance of Z is x it then follows that Z is false. But the Principal Principle requires that Z has positive credence x on this supposition.

Undermining problems arise because for Humeans, information about chance contains statistical information about the history of outcomes, and everyone agrees that such information is not in general admissible. To incorporate this fact, Lewis (1994) and Hall (1994) (independently) suggested a revised version of the Principal Principle, which I state in a form due to Hall (2004): if Cr is a rational initial credence function, $Ch_t = f$ the proposition that the chance function at t is f, and E arbitrary information, then

$$Cr(A/Ch_t = f \wedge E) = f(A/Ch_t = f \wedge E),$$

provided $Cr(Ch_t = f \wedge E) > 0$. Roughly speaking, the idea is to treat chance as an expert, so that conditional on the information that $Ch_t = f \wedge E$ one should align one's beliefs with the expert's opinion *conditioned on that same information*. Hall (1994; 2004) argues that the new formulation is preferable to the old one on independent grounds, even if (like Hall) one rejects Humeanism about chance.

The revised Principle avoids the undermining problems. In the case of the undermining future Z, $Ch_t = f$ is logically incompatible with Z given $f(Z) < 1$, in which case $Cr(Z/Ch_t = f) = f(Z/Ch_t = f) = 0$, as desired. More illuminating is perhaps the coin toss example, because it is easier to work through in detail. Let f_1 be the chance function according to T_1, and consider the binary sequences of length 10 compatible with the hypothesis that T_1 is the best system. Plausibly, these sequences all have five 0s and five 1s in a more or less random-looking distribution. Let H_1 be the set of these sequences. If the credence function Cr assigns equal probability to the members of H_1, it follows that $Cr(\omega/Ch_t = f_1) = f_1(\omega/Ch_t = f_1)$ for every sequence ω. This is because $Ch_t = f_1$ rules out all sequences outside H_1, and both f_1 and Cr are uniform within H_1.

A potential downside of the revised Principle, apart from its unintuitive complexity, is that it requires chance to be defined over a very rich set of propositions, including propositions of the form $Ch_t = f \wedge E$. Some advocates of the best system account have therefore preferred to stick with a weakened form of the old Principle. On this view, the old Principle usually holds to a good approximation and occasionally breaks down for artificial propositions like Z or in artificially simple scenarios like that of the coin tosses (see Hoefer 2007, Schwarz 2014).

Having defused the Big Bad Bug, Humeans are wont to turn the tables. Granted, Humean accounts are in tension with simple formulations of the chance–credence link, but non-Humean accounts seem to render any such link entirely unintelligble (see e.g.

van Fraassen 1989, pp. 80–86, Lewis 1986b, pp. xvf., Lewis 1994, p. 239, Loewer 2004, p. 1120). Suppose, following advocates of primitive propensities, that there is a basic physical quantity besides spin, mass, charge, etc. which maps propositions (or pairs of propositions) to numbers. Across logical space, the values of this quantity are independent of other facts about the world: in some worlds, high values generally go with true propositions, in others with false ones, and so on. Such a quantity might well exist; indeed, there might be many of them, assigning different numbers to propositions, all independent of their truth. But then how could it be a basic constraint of rationality to believe propositions to the extent that a particular one of these quantities assigns them high numbers? Why should credence follow this quantity rather than one of the others? Why take high numbers as indications of truth rather than low ones? Note that we have absolutely no evidence that our world is one where high numbers tend to go with true propositions, for we have no direct way to inspect this primitive quantity; physicists estimate chances by *using* the Principal Principle, inferring (for example) that frequently occurring outcomes must have high chance.

By contrast, if chance is a Humean quantity reflecting patterns in the distribution of actual outcomes, it is not at all mysterious that information about chance should constrain rational belief about outcomes. To be sure, it is not immediately obvious that this constraint takes the specific form of the Principal Principle (suitably weakened or revised to account for undermining), and some anti-Humeans have expressed skepticism that Humean accounts can really explain the Principle (see Black 1998, Strevens 1999, Hall 2004). But we have already seen a simple illustration of how this might go. In the coin toss example above, we have *derived* (central instances of) Hall's revised Principal Principle from the assumption that equal initial credence goes to random-looking sequences with a fixed ratio of 0s and 1s. This is a plausible constraint on rational credence, independent of the interpretation of chance. Observations of this kind are generalised in Schwarz (2014) (see also Loewer 2004, pp. 1122f., Hoefer 2007, sec.5 for related arguments).

20.6 Anti-Humean Intuitions

For Humeans, the history of occurrent events in a world is metaphysically prior to laws and chances. Anti-Humeans, on the other hand, often hold that laws and chances "produce", "guide", or "govern" the history of events and therefore must have metaphysical priority. As an objection to Humeanism, this line of thought has proved rather elusive, as it is not clear what the metaphors of production, guidance or governance are supposed to express.

Anti-Humeans sometimes try to cash out the objection by pointing at links between laws and chance on the one hand and explanation, counterfactuals, or causation on the other. From an anti-Humean perspective, possible worlds without primitive nomic features are lawless worlds in which any regularity in the history of events is a mere coincidence; hence there can be no explanation of why (in such a world) a glass shatters when hit by a rock, and nothing definite can be said about what would have happened if, counterfactually, a glass had been hit by a rock. This is an alien and unappealing picture. But of course it is not at all the Humean picture. For example, on Lewis's Humean account of counterfactuals, causation and explanation (see Lewis 1986b), Humean laws do support counterfactuals, are involved

in causation, and figure in causal explanations. What is revealed by the present objection is merely that Humeanism is a package deal: Humeans about laws and chance had better not accept anti-Humean accounts of counterfactuals, causation, explanation, etc.

Arguably the best way to motivate the anti-Humean position is to emphasize the intuitive gap between the *is* of occurrent events and the *ought* of laws and chance. Lange (2000, pp. 48–51) describes a possible world in which there is a single particle, traveling at constant velocity. Intuitively, such a world could be governed by Newton's laws, but also by various alternatives; the actual events are too sparse to settle the question. This kind of thought experiment is especially powerful in the case of chance (see e.g. Tooley 1987, pp. 42–7). Imagine a world with nothing but a thousand "coin toss" events: 517 heads and 483 tails. Does it follow that the chance of heads on each toss is exactly 0.517? Intuitively not; the sequence is quite compatible with the chance being 0.5 or 0.6 or even 0.9. But then the chances are not determined by the actual outcomes.

To suggest otherwise seems to conflate improbability with impossibility. If the chance of heads is 0.9, getting 517 heads in 1000 tosses is improbable; but improbable things can happen. In general, the hypothesis that a proposition A has positive chance should always be compatible with A being true. This is a version of what Bigelow et al. (1993) call the *Basic Chance Principle*. It is also a special case of Lewis's original Principal Principle: in the absence of inadmissible evidence, one should assign positive credence to A given that A has positive chance. On the best system account, by contrast, world histories with comparatively low chance can be ruled out *a priori*!

Humeans might try to escape these objections by denying the move from conceivability to metaphysical possibility, appealing to Kripke (1980). This has not been a popular response. Friends of the best system account generally take the account to provide an analysis or explication of the concept of chance, rather than an empirical hypothesis about its essence. In fact, Humeans are generally skeptical about essences and other substantive kinds of necessity and possibility. The usual reaction is therefore to bite the bullet and concede that the best system account is incompatible with some intuitions about chance. This is sometimes combined with an allegation that the intuitions come from dubious "theological" sources and hence should not carry much weight in the first place (see e.g. Loewer 1997, Beebee 2000). One might also argue that the intuitions arise from uncritical applications of the Principal Principle. But one need not be so dismissive. One can accept that the anti-Humean intuitions are part of our ordinary conception (or conceptions) of chance. Humean accounts cannot satisfy this aspect of the chance role. But sometimes perfection is impossible. Humean accounts at least satisfy many other aspects of the chance role, including those that actually matter to science – and they do so precisely because they deny the independence of chances from actual outcomes.

20.7 Probability in Science

When philosophers talk about chance, they mostly have in mind fundamental dynamical chance in the sense of Section 20.4: a forward-looking probabilistic quantity that figures in dynamical laws of fundamental physics. It is doubtful whether this kind of chance exists in

worlds like ours. The dynamical law of standard quantum physics, the Schrödinger equation, is deterministic. In general, the formalism of standard quantum mechanics does not seem to involve probabilities at all. (Wave function intensities are sometimes called probabilities, but interference effects preclude any direct interpretation as probabilities of underlying physical states.) Probabilities do figure in some rivals to standard quantum mechanics, notably in de Broglie–Bohm mechanics and the "GRW" theory of Ghirardi et al. (1986) and Pearle (1989). The probabilities in GRW even look like fundamental dynamical chances, but those in de Broglie–Bohm mechanics do not, since they are not dynamical.

This highlights an advantage of the best system account that we have briefly touched upon in Section 20.4: unlike, e.g., the primitive propensity account, the best system account can accommodate objective physical probabilities that do not take the form of fundamental dynamical chances. Many advocates of the best system approach argue that it can serve as a general analysis of objective empirical probabilities, including, for example, the probabilities in de Broglie–Bohm mechanics.

Outside fundamental physics, probabilities are found all across science – from statistical mechanics to theories of genetic mutation and evolution. Here, too, the probabilities rarely look like propensities. In statistical mechanics, for example, they primarily pertain to different microphysical realizations of thermodynamic states; in coalescence theory, they are even "backward-looking", pertaining to the time in the past at which genes in the present generation had their latest common ancestor. On the other hand, these probabilities satisfy many of the conditions discussed in Sections 20.4 and 20.5, such as the connection to frequencies, randomness, and rational credence. This suggests that they might be interpreted along the lines of the best system account.

The most ambitious and detailed development of this idea is the Albert-Loewer account of statistical mechanics (see Albert 2000, Loewer 2007a, Loewer 2012). Albert and Loewer argue that the complete laws of physics include, besides the dynamical microphysical laws (assumed to be deterministic), a "Past Hypothesis" to the effect that the universe started in a certain thermodynamic state M_0 of low entropy, as well as a uniform probability distribution PROB over microstates realizing M_0. These probabilities, they argue, are best understood as chances in the sense of the best system account. The basic idea is that the package of microphysical laws, Past Hypothesis and PROB provides the most elegant and informative summary of the events in our world. The microphysical laws by themselves entail almost nothing about the behaviour of macroscopic objects, unless one happens to know their precise microstate. Adding the probability distribution over initial states, Loewer (2007a, p. 305) argues, "results in a system that is only a little less simple but is vastly more informative".

This proposal has been attacked from different directions. Setting aside worries about its physical tenability (see, e.g., Earman 2006), there are doubts about its application of the best system approach. By the standards of Section 20.3, adding PROB and the Past Hypothesis to the micro-laws would not create a system "that is only a little less simple", since neither addition can be expressed in the fundamental language of microphysics (or if so, only as an infinite disjunction). Loewer therefore remarks that theories may include principles not only in the fundamental language, but also in the language of thermodynamics (Loewer 2007a, p. 305, fn.23). One may wonder what general rule this comes from, and whether the system composed of micro-laws, Past Hypothesis, and PROB would really come out best. The mere fact that it is better than the micro-laws alone hardly settles that matter. Other worries arise from the fact that Albert and Loewer want the Past Hypothesis to be a *law*, which goes

against the usual conception of dynamical laws and non-lawful boundary conditions. (See, e.g., Frisch (2011) and Winsberg (2008; 2013) for concerns along these lines.)

Perhaps the most controversial aspect of the Albert-Loewer proposal is the idea that all probabilities in the special sciences are instances of PROB. Many philosophers want to grant more autonomy to the special sciences. This is easily achieved in the framework of the best system account: the laws of genetic mutation, for example, can be understood as best systematizations of facts *about genetic mutation*; the basic vocabulary here is not the language of physically fundamental properties, but a suitable language of chemical and biological kinds (see Schrenk 2007, Cohen and Callender 2009; 2010). A variation of this view assumes a single best system for all of science, but allows it to have autonomous parts for different branches of science. The addition of special science laws is then motivated by the fact that they provide systematizations of useful facts about the world that cannot be discerned in the microphysical patterns, or not without impossible computational effort (see Hoefer 2007, Frigg and Hoefer 2010, Frisch 2011).

These applications of the best system account to probabilities in special science demand some modifications to the rules of Section 20.3. Most obvious are issues arising for the specification of the relevant language. In addition, there are different ways of taking into account the fact that most special science laws are only *ceteris paribus* laws. It is also plausible that new virtues apply to special science theories, such as integration with other areas of science. These details still remain to be spelled out, in what promises to be a fruitful area of future research.

References

Albert, D. (2000) *Time and Chance*. Cambridge MA: Harvard University, Press.

Arntzenius, F. and Hall, N. (2003) On What We Know About Chance. *British Journal for the Philosophy of Science*. 54. 2.

Beebee, H. (2000) The Non-Governing Conception of Laws of Nature. *Philosophy and Phenomenological Research*. 61. pp. 571–94.

Bernstein, A. R. and Wattenberg, F. (1969) Non-Standard Measure Theory. In Luxemburg, W. (ed.) *Applications of Model Theory of Algebra, Analysis, and Probability*. New York, NY: Holt, Rinehart and Winston.

Bigelow, J., Collins, J., and Pargetter, R. (1993) The Big Bad Bug: What are the Humean's Chances? *British Journal for the Philosophy of Science*. 44. pp. 443–62.

Black, R. (1998) Chance, Credence and the Principal Principle. *British Journal for the Philosophy of Science*. 49. pp. 371–85.

Cohen, J. and Callender, C. (2009) A Better Best System Account of Lawhood. *Philosophical Studies*. 145. 1. pp. 1–34.

Cohen, J. and Callender, C. (2010) Special Sciences, Conspiracy and the Better Best System Account of Lawhood. *Erkenntnis*. 73. pp. 427–47.

Earman, J. (1984) Laws of Nature: The Empiricist Challenge. In Bogdan, R. (ed.) *D. M. Armstrong*. Dordrecht: Reidel.

Earman, J. 1986) *A Primer on Determinism*. Dordrecht: Reidel.

Earman, J. (2006) The 'Past Hypothesis': Not Even False. *Studies in History and Philosophy of Science*. 37. pp. 399–430.

Earman, J. and Roberts, J. T. (2005) Contact with the Nomic: A Challenge for Deniers of Humean Supervenience about Laws of Nature. Part I: Humean Supervenience. *Philosophy and Phenomenological Research*. 71. pp. 1–22.

Elga, A. (2004) Infinitesimal Chances and the Laws of Nature. In Jackson, F. and Priest, G. (eds.) *Lewisian Themes: The Philosophy of David K. Lewis*. pp. 68–77. Oxford: Oxford University Press.

Frigg, R. and Hoefer, C. (2010) Determinism and Chance From a Humean Perspective. In Stadler, F. et al. (eds.) *The Present Situation in the Philosophy of Science.* pp. 351–72. Dordrecht: Springer.

Frisch, M. (2011) From Arbuthnot to Boltzmann: The Past Hypothesis, the Best System, and the Special Sciences. *Philosophy of Science*. 78. pp. 1001–11.

Gaifman, H. and Snir, M. (1982) Probabilities Over Rich Languages, Testing and Randomness. *Journal of Symbolic Logic*. 47. 3. pp. 495–548.

Ghirardi, G., Rimini, A., and Weber, T. (1986) Unified Dynamics for Micro and Macro Systems. *Physical Review D*. 34. pp. 470–91.

Giere, R. N. (1973) Objective Single Case Probabilities and the Foundations of Statistic. In Suppes, P. et al. (eds.) *Logic, Methodology and the Philosophy of Science IV.* pp. 467–83. Amsterdam: North Holland.

Hájek, A. (1997) "Mises redux" – redux: Fifteen Arguments Against Finite Frequentism. *Erkenntnis*. 45. 2-3. pp. 209–27.

Hall, N. (1994) Correcting the Guide to Objective Chance. *Mind*. 103. pp. 505–17.

Hall, N. (2004) Two Mistakes about Credence and Chance. *Australasian Journal of Philosophy*. 82. pp. 93–111.

Hoefer, C. (2007) The Third Way on Objective Probability: A Skeptic's Guide to Objective Chance. *Mind*. 116. pp. 549–96.

Joyce, J. (2009) Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Huber, F. and Schmidt-Petri, C. (eds.) *Degrees of Belief.* 263–300. Berlin: Springer.

Kripke, S. (1980) *Naming and Necessity*. Oxford: Blackwell.

Lange, M. (2000) *Natural Laws in Scientific Practice*. Oxford: Oxford University Press.

Levi, I. (1990) Chance. *Philosophical Topics*. 18. pp. 117–49.

Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.

Lewis, D. (1980) A Subjectivist's Guide to Objective Chance. In Jeffrey, R. (ed.) *Studies in Inductive Logic and Probability.* Vol. 2. Berkeley: University of California Press.

Lewis, D. (1983) New Work for a Theory of Universals. *Australasian Journal of Philosophy*. 61. pp. 343–77.

Lewis, D. (1986a) *On the Plurality of Worlds*. Malden, MA: Blackwell.

Lewis, D. (1986b) *Philosophical Papers II*. New York, NY and Oxford: Oxford University Press.

Lewis, D. (1994) Humean Supervenience Debugged. *Mind*. 103. pp. 473–90.

Loewer, B. (1997) Humean Supervenience. *Philosophical Topics*. 24. pp. 101–26.

Loewer, B. (2001) Determinism and Chance. *Studies in History and Philosophy of Science Part B*. 32. 4. pp. 609–20.

Loewer, B. (2004) David Lewis's Humean Theory of Objective Chance. *Philosophy of Science*. 71. pp. 1115–25.

Loewer, B. (2007a) Counterfactuals and the Second Law. In Price, H. and Corry, R. (eds.) *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited.* pp. 293–326, New York, NY: Oxford University Press.

Loewer, B. (2007b) Laws and Natural Properties. *Philosophical Topics*. 35. pp. 313–28.

Loewer, B. (2012) Two accounts of laws and time. *Philosophical Studies*. 160. pp. 115–37.
Mellor, D. H. (1971) *The Matter of Chance*. Cambridge: Cambridge University Press.
Mill, J. S. (1843) *A System of Logic*. London: Parker.
Myung, I. J., Balasubramanian, V., and Pitt, M. A. (2000) Counting Probability Distributions: Differential Geometry and Model Selection. *Proceedings of the National Academy of Sciences*. 97. 21. pp. 11170–75.
Pearle, P. (1989) Combining Stochastic Dynamical State-vector Reduction with Spontaneous Localization. *Physical Review*. 39A. pp. 2277–89.
Popper, K. (1982) *Quantum Theory and the Schism in Physics*. Totowa, NJ: Rowman and Littlefield.
Ramsey, F. P. (1978) Universals of Law and of Fact. In *Foundations*. London: Routledge & Kegan Paul.
Rissanen, J. (1978) Modeling by the shortest data description. *Atomatic*. 14. pp. 465–71.
Rissanen, J. (1989) *Stochastic Complexity in Statistical Inquiry Theory*. River Edge, NJ: World Scientific Publishing Co.
Schaffer, J. (2007) Deterministic Chance? *British Journal for the Philosophy of Science*. 58. pp. 113–40.
Schrenk, M. A. (2007) A Lewisian Theory for Special Science Laws. In Bohse, H. and Walter, S. (eds.) *Philosophie: Grundlagen und Anwendungen. Ausgewählte Beiträge aus den Sektionen der GAP 6.* Paderborn: Mentis.
Schwarz, W. (2014) Proving the Principal Principle. In Wilson, A. (ed.) *Chance and Temporal Asymmetry.* pp. 81–99. Oxford: Oxford University Press.
Strevens, M. (1999) Objective Probability as a Guide to the World. *Philosophical Studies*. 95. pp. 243–75.
Tooley, M. (1987) *Causation: A Realist Approach*. Oxford: Oxford University Press.
van Fraassen, B. C. (1989) *Laws and Symmetry*. Oxford: Clarendon Press.
Winsberg, E. (2008) Laws and Chances in Statistical Mechanics. *Studies in History and Philosophy of Modern Physics*. 39. pp. 872–88.
Winsberg, E. (2013) The Metaphysical Foundations of Statistical Mechanics: On the Status of PROB and PH. In Loewer, B., Winsberg, E. and Weslake, B. (eds.) *[Currently untitled],* Cambridge, MA: Harvard University Press.

CHAPTER 21

PROBABILITY AND RANDOMNESS

ANTONY EAGLE

Random outcomes, in ordinary parlance, are those that occur haphazardly, unpredictably, or by chance. Even without further clarification, these glosses suggest an interesting connection between randomness and **probability**, in some of its guises. But we need to be more precise, about both probability and randomness, to understand the relationship between the two subjects of our title.

It is a commonplace that there are many sorts of probability; and of each, we may ask after its connection with randomness. There is little systematic to say about randomness and **credence**; even rational degrees of belief may reach certainty about a random outcome, and remain at uncertainty about a non-random one. More of substance can be said about connections between randomness and **evidential probability** – particularly according to Solomonoff's version of objective Bayesianism, which uses Kolmogorov complexity (Section 21.2.2) to define a privileged prior **algorithmic probability** (Solomonoff 1964; see also Li and Vitanyi 2008; Rathmanner and Hutter 2011).

For reasons of both concision and focus, in the present chapter I set those issues aside, to concentrate on randomness and **physical probability**, or **chance**: probability as a physical feature of certain worldly processes.[1] A number of philosophers have proposed an intimate connection between randomness and chance, perhaps even amounting to a reduction of one to the other. I explore, with mostly negative results, the prospects for such views; and discuss some weaker but still interesting ways in which randomness bears on chance. I begin by clarifying and distinguishing a number of kinds of randomness.

21.1 Different Kinds of Randomness: Product and Process

The paradigm sort of case that might involve randomness is a series of tosses of a fair coin, or similar chance device. A **typical** example of such a series will be a disorderly and patternless

[1] For more on the nature of chance, see the chapters by Frigg, Gillies, La Caze, and Schwarz (22), (19), (16), and (20) in this volume.

sequence of outcomes (Heads and Tails, or 0s and 1s), because it will have been produced by a genuinely chancy process.

However, it is at least conceivable that a chancy process could produce an orderly series of outcomes – a lucky series of coin tosses may land all heads. It is similarly conceivable that a disorderly series of outcomes could have been produced by a non-chancy process. It seems arbitrary to believe that, in the more typical case, it is really just one of these features that is responsible for the presence of randomness. Hence we might find it useful to regiment our terminology, distinguishing different sorts of randomness in virtue of the different features that can prompt ascriptions of randomness.

A sequence of 100 consecutive Heads is *prima facie* not random, even if produced by fair tosses of a fair coin. Yet a sequence of 100 mixed Heads and Tails from the same fair coin with no discernible pattern is *prima facie* random. The difference here in the *prima facie* appearances cannot be grounded in the underlying chances, which are the same in both cases (each has chance $1/2^{100}$). The appearance is grounded rather in the intrinsic disorderliness (or otherwise) of the outcome sequences themselves. Let's say that a **random product** is a series of outcomes of some repeated process that is disorderly and irregular, regardless of how it was produced. Moreover, let's say that a **random process** is one involving genuine chance, in which some of the possible outcomes of the process have non-trivial objective chances of coming to pass. There is precedent for this regimentation, assuming that 'probabilistic laws' support objective chances:

> I group *random* with stochastic or chancy, taking a random process to be one which does not operate wholly capriciously or haphazardly but in accord with stochastic or probabilistic laws.
>
> (Earman 1986: p. 137; my italics)

This proposed regimentation diverges to a certain extent from intuition, in some cases where the chances are nearly 1 or nearly 0. An exercise of a quite reliable skill, such as catching a ball, can be a random process, if there is some genuine chance of failing to catch the ball. It is awkward to characterize the occurrence of the – entirely expected – outcome of such a process as random. In the case of human action, perhaps our judgments are being confounded by the fact that, while my catching a ball is partly a matter of chance, it is not solely due to chance. So instead we should say: a process is random if at least some of its outcomes are *purely* a matter of chance – they happen not just 'in accord with' probabilistic laws, but are adequately explained by citing probabilistic laws.

Even refined in this way, the regimentation classifies some highly probable outcomes as random – any event which is best explained by citing a chancy law of nature. Suppose we consider the outcomes of this repeated chance process: roll two fair dice, record '1' if double six comes up, and '0' otherwise. In the long run, the frequency of 1s in the outcome sequence will be (close to) 1/36. Because of this, the outcome sequence will be quite orderly: it will be almost all 0s, with a few 1s scattered here and there. Yet this process is purely chancy (assuming that dice rolls are), and so is random. Ordinary intuition appears to be split on such cases. On the one hand, the outcomes are quite regular, and even predictable (if you always bet on '0', you'd be almost always right). On the other, the best way to form an opinion about a given outcome is simply to set your credence in line with the chances, and the best explanation of why some outcome came up in a particular case just cites the chances. Once we make the distinction between product and process randomness, we

can vindicate both these intuitive judgments: this scenario involves a random process, but produces something non-random. Those who would refuse to call biased chance processes 'random' are – I venture – letting views about the randomness of the product drive their opinion of randomness of the process. We have a more useful taxonomy if we keep these two categories apart.

On my suggested regimentation, there is a definitional link between process randomness and physical probability. The relationship between product randomness and physical probability is less clear. The typical outcome of a repeated random process will be a random product; but it is conceivable both that some random products are not produced by chance, and, that some repeated trials of random processes don't yield random products. Having regimented our vocabulary so as to be able to make such distinctions, it is clear that we can neither **define** randomness of a product in terms of randomness of the process generating it, nor vice versa.

But for all that, it may be that we can **reduce** one to the other, finding some metaphysical account which either grounds the randomness of a sequence in the randomness of some process, or the converse. This converse reduction of process randomness to product randomness is the characteristic idea behind the **frequentist** account of chance (von Mises 1957; La Caze, this volume (16)). Von Mises claims, in effect, that a process is chancy iff it would produce a random sequence of outcomes which exhibit stable frequencies; those frequencies are to be identified with the chances of those outcomes. Frequentists often proposed a definitional link between chance and frequencies. They needn't have; they need only claim that frequencies in random sequences **actually realize** the chances.

I begin the rest of this chapter by exploring proposals in the vicinity of frequentism in this respect: proposals that attempt to ground process randomness – at least in part – in facts about product randomness. That task involves giving a theory of product randomness that makes no reference to chance or probability, a theory that characterises disorderliness of a sequence directly. In Section 21.2, I discuss first von Mises' attempt to provide such a theory, then related theories due to Martin-Löf and Kolmogorov. In Section 21.3, I examine whether there is any viable prospect of using randomness of sequences to ground facts about chance. My conclusions are largely negative. At the end of Section 21.3, I discuss the weaker claim that product random sequences can be good **evidence** for chances. Above, I noted our ability to distinguish orderly and disorderly sequences, independently of the randomness or otherwise of the underlying process. If we can come up with a sensible account of product randomness, we may be able to use the existence of a product random sequence as evidence for the existence of a random process. It seems that if the outcome sequence is product random, it will be unpredictable; and accordingly it will be epistemically irrational to seek certainty in advance concerning future outcomes. In such a situation, probabilistic theories look particularly attractive, providing reliable information about the future when certainty cannot be had. Given familiar difficulties concerning the epistemology of chance, the possibility that product randomness is evidentially significant for chance is a good reason to explore product randomnesmore deeply.

The distinction between process and product randomness doesn't exhaust the theoretically useful categories in this vicinity. Typically, random processes yield **unpredictable outcomes**, in the sense that gathering additional information concerning already observed outcomes doesn't generally improve predictive accuracy compared to making estimates

based solely on the underlying chances. Likewise, a random product involves unpredictable outcomes, in that information about earlier outcomes in the sequence doesn't provide information about later outcomes in the sequence – for if it did, that would be a pattern (perhaps somewhat subtle) in what is supposed to be a disorderly sequence. (We might be mystified by the process involved, but a sequence of coin tosses which reliably came up heads every prime-numbered toss would be at least partially predictable.) So both product and process randomness tend to lead to unpredictability.[2] But the converse isn't true. A process may be unpredictable because we can't work out how to model it, even though it is perfectly deterministic and non-chancy. (Some examples like this are discussed in Section 21.3.2 below.) Likewise, a product may be unpredictable, even though there is an underlying pattern or order, because that pattern is too difficult to discern.

Elsewhere, I've argued that it would be theoretically elegant to identify randomness with unpredictability, on the grounds that only such a liberal identification would capture all cases of randomness (Eagle 2005). My basic argument was that product and process randomness are each too narrow to apply to everything characterized as random by the various special sciences. The notion of unpredictability I offered was, however, explicated just in terms of credence, and didn't say much about how these 'random' phenomena justify our subjective probability assignments. In focussing on the connections between randomness and physical probability in this chapter, I will discuss some outcomes that are unpredictable because they are process and/or product random. But I will neglect the class of unpredictable outcomes more generally, because there is little both systematic and true that can be said about the relationship between unpredictability and physical probability.

I also largely set aside another interesting notion in the vicinity: **pseudorandomness**. Like product randomness, pseudorandomness is a property of sequences. Informally, a sequence is pseudorandom if it *looks* disorderly. More precisely, a sequence will be pseudorandom if it passes sufficiently good tests for randomness, in that 'no efficient algorithm can distinguish [pseudorandom] output from a truly random sequence' (Vadhan 2012: §1.1). So a pseudorandom sequence is not product random, because there exists some pattern underlying the sequence. But that pattern is not able to be efficiently exploited – so for all practical purposes, the sequence is as good as random. The interest of pseudorandomness lies in the question as to whether the use of pseudorandom sequences might *replace* the use of random sequences in many applications, for example in cryptography. But since pseudorandom sequences are, by construction, neither produced by chance nor genuinely patternless, they are not random in the core sense that concerns us here.

[2] The same sorts of worries about the randomness of outcomes of very high and very low chance arise again here. It is very plausible that if your best estimate of whether an outcome will happen is no better than chance, then you are not able to predict that outcome. But nearly-extreme-chance outcomes certainly appear predictable; and indeed, it is often possible to know propositions that have high but not extreme chance (Hawthorne and Lasonen-Aarnio 2009). So it can turn out (i) that the best credence to have in a high chance outcome is just equal to its chance, which makes that outcome unpredictable; but (ii) that for all practical purposes simply assuming that it will occur is the best strategy; and (iii) it may even be that, though unpredictable, one can know the outcome will occur. This is one more nice illustration of the sometimes uncomfortable fit between traditional and Bayesian epistemology.

21.2 Algorithmic Randomness

Laplace observed that, when tossing a coin,

> if heads comes up a hundred times in a row, then this appears to us extraordinary, because the almost infinite number of combinations that can arise in a hundred throws are divided in regular sequences, or those in which we observe a rule that is easy to grasp, and in irregular sequences, that are incomparably more numerous.
>
> (Laplace 1951: pp. 16–17)

Modern theories of product randomness take up two themes from Laplace – first, that random sequences are not governed by a rule (whether or not 'easy to grasp'); second, that random sequences are more numerous than non-random ones. Following the literature on **algorithmic randomness**, we will restrict our attention to binary sequences of outcomes, like sequences of outcomes of tosses of a fair coin. The mathematical literature on random sequences is extensive (for further details, see Eagle 2014; Dasgupta 2011; Li and Vitanyi 2008; Nies 2009; and references therein).

As mentioned above, von Mises' version of frequentism required, to avoid circularity, a characterization of product randomness that did not involve an antecedently given probability function. The characterization he offered applies only to infinite sequences, and it is convenient to begin by considering these. In fact, we will confine our attention to infinite binary sequences (sequences of 0s and 1s), the sort that might model an infinite series of coin tosses. The set of all infinite binary sequences we call the **Cantor space**.

21.2.1 Infinite Random Sequences: Von Mises' Approach

Von Mises' approach illustrates one of the Laplacean themes: the **unruliness** of random sequences. Considering his example of regular milestones spaced along a road, of which a probabilistic account would be inappropriate, he claims that

> the essential difference between the sequence of the results obtained by casting dice and the regular sequence of large and small milestones consists in the possibility of devising a method of *selecting the elements* so as to produce a fundamental change in the relative frequencies.
>
> (von Mises 1957: p. 24)

The rule governing a non-random sequence can be exploited to construct a method to select a biased partial sequence from the original sequence: for example, the rule *select every element which occurs after nine small stones* can be used to select a partial sequence of the milestones in which the limit relative frequency of large stones is 1, even though we never use that attribute directly in selecting outcomes for the subsequence. However, in a random sequence, claims von Mises, we are unable to use such a method – the only way to obtain a biased partial sequence of outcomes is to choose outcomes based on their attributes directly.

Accordingly, he offers a characterization of random sequences as those in which the **limiting frequencies are invariant in partial sequences**:

> these limiting values must remain the same in all partial sequences which may be selected from the original one in an arbitrary way. Of course, only such partial sequences can be taken into consideration as can be extended indefinitely, in the same way as the original sequence itself. Examples of this kind are, for instance, the partial sequences formed by all odd members of the original sequence, or by all members for which the place number in the sequence is the square of an integer, or a prime number, or a number selected according to some other rule, whatever it may be. The only essential condition is that the question whether or not a certain member of the original sequence belongs to the selected partial sequence should be settled *independently of the result* of the corresponding observation, i.e., before anything is known about this result. We shall call a selection of this kind a *place selection*. The limiting values of the relative frequencies in a collective must be independent of all possible place selections. By place selection we mean the selection of a partial sequence in such a way that we decide whether an element should or should not be included without making use of the attribute of the element, i.e., the result of our game of chance.
>
> (von Mises 1957: pp. 24–5)

It is trivial to observe that – without further restrictions on what kind of place selections are admissible – there will be no random sequences, because no limit frequencies are invariant under arbitrary place selections. For if a place selection is just any function f from natural numbers into $\{0, 1\}$,[3] then there will be a place selection that selects a biased subsequence, since

> any increasing sequence of natural numbers $n_1 < n_2 < \cdots$ defines a corresponding selection rule, ... given an arbitrary sequence of 0s and 1s ... there is among the selection rules the one which selects the 1s of the given sequence, so the limit frequency is changed
>
> (Martin-Löf 1969a: p. 27)

Von Mises does implicitly place restrictions on appropriate place selections. The empirical basis for von Mises' claim that there exist sequences which are appropriate for probabilistic treatment lies in the **principle of the impossibility of a gambling system** (von Mises 1957: p. 25): in some systems, there is no recipe for deciding when to bet on outcomes that ensures more successes than chance alone. (The 'gambler's fallacy' is an unusually simple gambling system, but its lack of success is entirely representative.) The 'recipe' *bet only on 1s* does exist, in some abstract mathematical sense, but it is not one that can be followed by anyone. So von Mises' implicit restriction is to those place selections that can be genuinely implemented by a prospective gambler, making use only of information in the gambler's possession (information about previous outcomes). Von Mises made his original proposal before Turing and others made the notion of an **algorithm** precise, but it was very natural to retrospectively read into his work a restriction to **computable** place selections:

> To a player who would beat the wheel at roulette a system is unusable which corresponds to a mathematical function known to exist but not given by explicit definition; and even the explicit definition is of no use unless it provides a means of calculating the particular values of the function. ... Thus a [place selection] should be represented mathematically, not as a

[3] I.e., if $f(n) = 1$, then and only then select the nth member of the original sequence to belong to the partial sequence.

> function, or even as a definition of a function, but as an effective algorithm for the calculation of the values of a function.
>
> (Church 1940: p. 133)

Where x is some infinite sequence, let '$x \upharpoonright i$' denote the initial segment of x of length i. Church proposes that an implementable place selection is an **effectively computable function** $\phi : (x \upharpoonright (i-1), i) \rightarrowtail \{0,1\}$ that takes the value 1 infinitely many times (so always selects an infinite partial sequence from an infinite sequence). Let $x[\varphi]$ be the infinite partial sequence consisting of any x_j such that $\varphi\left(x \upharpoonright \left(j-1\right), j\right) = 1$. According to Church's reconstruction of von Mises, a sequence is random iff for every effectively computable place selection φ, the limiting frequency of every outcome in $x[\varphi]$ equals the limiting frequency of that outcome in x. More informally: there exists no effective method for choosing a subsequence of a random sequence that is biased with respect to the original frequencies. The random sequences in this sense are immune to gambling systems: there are no **followable rules** that enable even fallible and unreliable prediction of patterns of outcomes in a random sequence, or any partial sequence derived from it, except strictly in accordance with the statistical predictions of its probabilistic laws.

The question of whether there are any random sequences in this revised sense still needs to be addressed. As the number of effectively computable functions is much smaller than the number of arbitrary functions, it is easier to satisfy the requirement of invariance of limit frequencies under all computable place selections. In fact, there do exist infinite random sequences in this refined von Mises/Church sense. Indeed, **almost all** infinite binary sequences have the feature of being frequency-invariant under computable place selections. More precisely, the set of von Mises/Church random sequences forms a **measure one** subset of the Cantor space. This follows quickly from a result due to Wald (1938), that any **countable** set of arbitrary place selections (whether computable or not) defines a set of von Mises random sequences which is measure one, given that there are only countably many computable place selections. Pursuing the unruly behaviour of random sequences has now led us to the other aspect of random sequences Laplace noted: the von Mises/Church random sequences are much more numerous than the even partially rule-governed sequences.

21.2.2 Infinite Random Sequences: The Typicality Approach

What if we had started instead from Laplace's other claim: that irregular sequences are more numerous than regular sequences? Suppose we know that a fair coin is to be tossed repeatedly; what sort of outcome sequence should we expect? Well, we ought to expect that heads should occur about as often as tails; that strings of heads and tails of equal length should occur about as frequently as each other; and so on. These are features that are **typical** of sequences generated by this sort of chance setup. Even if we have very little confidence in any specific sequence of outcomes occurring, we should be confident that some typical sequence will occur. This confidence derives ultimately from the fact that typical sequences are much more common than atypical ones.

This last claim needs to be handled with some care. Given an uncountably infinite set, like the Cantor space, there are many ways of measuring the size of its subsets. We want the

typical sequences to form a measure one set of sequences, but *under which measure*? This is particularly important when we consider biased binary processes. The typical outcome of a series of flips of a coin biased to heads will be a decidedly atypical product of a series of fair coin flips. The standard way of approaching this issue is to let the probability function associated with the underlying binary process determine a measure over the Cantor set; thus typicality of a sequence is relative to a probability function (Martin-Löf 1966; Gaifman and Snir 1982).

Obviously, and in contrast to the von Mises/Church account, such approaches to randomness rely on antecedent probabilities, and thus are generally unsuited to the reductive project we're considering. But there is a special case, where we define a measure not from an underlying chance, but from symmetries in the Cantor space itself. The measure thus produced is a very natural one over the Cantor space: the **Lebesgue measure** λ. The key idea is that the set of all sequences which begin with a given initial subsequence x_n of length n has Lebesgue measure $1/2^n$. (So the set of sequences beginning with a 1 has measure 1/2, the set of sequences beginning 010 has measure 1/8, etc.) Every particular infinite sequence $x \in X$ has Lebesgue measure zero (it must have measure smaller than any set of sequences sharing an initial subsequence with it, and $\lim_{n\to\infty} \frac{1}{2^n} = 0$). So indeed does any countable set of infinite binary sequences. The set of von Mises/Church random sequences has Lebesgue measure one, because its complement – the set of effectively computable binary sequences – is countable, connected as it is with the countable set of effectively computable functions, and thus has Lebesgue measure zero. The von Mises/Church random sequences are typical with respect to the Lebesgue measure over the Cantor space. (Henceforth, I will simply say they are ***typical***.)

A typical sequence has some **measure one property**: one that is shared by almost all the sequences. But there are lots of distinct measure one properties. For example, the property of having 0 and 1 occur equally frequently is a property of almost all sequences. It is also, intuitively, a property that a random sequence should have. Intuitively, a random sequence should have the stronger property of **Borel normality** (Borel 1909): every string of 1s and 0s of length n occurs equally often in x. (So 010 occurs as often as 101 and 111, etc.) Borel normality is also a measure one property. Indeed, every measure one property that has been explicitly considered is one that, intuitively, random sequences have. Generalizing from this, Ville (1939) proposed that a random sequence should be a member of *all* measure one sets of sequences – the random sequences are **typical in every respect**.

The typicality approach, in this crude form, cannot work. For each individual sequence has measure zero, so its **complement** with respect to the Cantor space has measure one and excludes it; so the intersection of *all* sets of measure one is the empty set. No sequence is typical in every respect.

This is reminiscent of the way in which consideration of arbitrary functions trivialized von Mises' original theory of place selections. Martin-Löf (1966) proposed an analogous solution: rather than considering arbitrary measure one sets, he suggests that the random sequences should satisfy all **effective** measure one properties. A set has Lebesgue measure zero if there is a sequence of sets which converge to it, such that the measures of the sets converge to zero. If that sequence of sets is **effectively computable** (so that there is a computable function which computes what each member of the sequence is), then the set has effective measure zero. A sequence is **Martin-Löf random** iff it is not a member of any

effective measure zero set – i.e., has all effective measure one properties. The Martin-Löf random sequences are **typical in every effectively determinable respect**. (More precisely: they are not effectively determinable to be atypical.) All the properties we have considered so far – Borel normality, non-computability, and the property of symmetric oscillation considered at the end of this section – are effective measure one.

Martin-Löf was led to this idea by considering **significance tests** in statistics. An experimental outcome prompts a hypothesis to be rejected at a significance level α if the outcome falls into a previously specified critical region. A sequence which falls into an effective measure zero set is one that would effectively yield a statistically significant result at arbitrarily high levels of significance. That is, it would prompt us to reject the null hypothesis that the sequence is random with arbitrarily high significance, and so must really be non-random. In this reformulation, Martin-Löf's proposal is that a sequence is random iff it passes all recursive statistical tests for randomness.

To show that there are Martin-Löf random sequences, Martin-Löf proves first that there is a **universal** significance test: there exists an effectively specifiable sequence of sets $U = U_1, \ldots$, such that for any other significance test $G = G_1, \ldots$, there exists a constant c such that for all i, $G_{i+c} \subseteq U_i$. If a sequence is rejected by some test at some significance level, it will also be rejected by U at some related significance level. Since the non-random sequences are those which, for any significance level, are rejected by some test, they will all be rejected by the universal test too. This universal test thus directly establishes that the set of Martin-Löf non-random sequences has effective measure zero; so Martin-Löf random sequences exist and collectively have effective measure one.

What is the relationship between von Mises' approach and the typicality approach? Let's begin by considering another way to think about random sequences. Each random sequence x corresponds to a **random walk** on a line, in which the walker begins at some starting point, and moves one unit left at stage i if the i-th element of x is 0, and moves to the right otherwise. The behaviour of this trajectory can be used to illustrate properties of random sequences. For example, the typical random walk ends up back at the starting point, since it will have moved left exactly as often as it will right. This corresponds to an infinite binary sequence having a limit relative frequency of 1/2 for 0 and 1.

Intuitively, a genuinely random walk should also cross the starting point often enough that it spends as much time, in the long run, on the left of its starting point as the right.[4] A walk that never spent any time to the left of the starting point has too much structure to be truly random. Happily for Martin-Löf's theory, this property of **symmetric oscillation** does have effective measure one (Dasgupta 2011: §3.4).

But Ville showed that there exist infinite binary sequences which are von Mises/Church random, and yet have biased initial segments: while the limit frequency of 1s in the sequence and all infinite subsequences is 1/2, the limit is approached 'from above', and the frequency of 1s in every finite initial subsequence is $\geq 1/2$ (Ville 1939; see also Lieb et al. 2006; van Lambalgen 1987). Such sequences have the same limit as the genuinely random

[4] Not only should a random sequence have a limit relative frequency of 1s equal to 1/2; it should approach this limit by having as many finite initial subsequences in which the relative frequency is below 1/2 as it has in which the relative frequency is above 1/2. The idea that random processes should spend equal amounts of time in all equally sized outcome states (the states here are *being on the left of the origin* and *being on the right*) is related to ergodicity and its relatives, discussed in Section 21.3.2 below.

sequences, but are not random in *how* they approach that limit. So there is a property that, intuitively, random sequences have (and that Martin-Löf random sequences have, in agreement with intuition), that some von Mises/Church random sequences lack. So while von Mises/Church randomness is necessary for Martin-Löf randomness, it is not sufficient (see also van Lambalgen 1987: §4).

21.2.3 Finite Random Sequences

A major difficulty facing Martin-Löf's proposal is that it cannot accommodate random finite sequences, or **strings**. Every string can be effectively produced by some Turing machine (even if not particularly *efficiently*), so no string can be even von Mises/Church random.

But while the contrast between sequences exhibiting effectively exploitable patterns and 'completely lawless' random sequences does not apply to finite sequences, a related contrast does apply: between those strings which can be efficiently described, and those cannot. This leads us to the idea that finite random sequences, like their infinite cousins, are patternless, or disorderly, and cannot be predicted (or exploited) by a simple rule.

Kolmogorov was the first to connect this idea of disorder with **incompressibility** (Kolmogorov 1963; Kolmogorov and Uspensky 1988). The informal idea is that a pattern in a string can lead to a short description of its contents, while disorderly strings cannot be described more effectively than by stating their contents. So while the string consisting of 1000 alternating 1s and 0s can be simply specified (I just specified it in nine words), some string consisting of 1000 1s and 0s with no pattern is most effectively presented by just listing the 1000 digits comprising it. This can be made more precise by thinking about **compressibility**.

If f is a **computable encoding function** from strings to strings, we say that a string δ is an **f-description** of a string σ iff $f(\sigma) = \delta$. A string σ is **compressed** by f if there is an f-description δ of σ such that $|\delta| < |\sigma|$, where '$|\varphi|$' denotes the length of the string φ. We may define the **f-complexity** of a string σ, $C_f(\sigma)$, as the length of the shortest string δ that f-describes σ. A string σ is **random relative to f** iff it is f-incompressible, that is, if the f-complexity of σ is roughly equal to $|\sigma|$.

This gives us randomness relative to a fixed algorithm f. It is in the spirit of the Laplacean ideas with which we began; since strings random relative to f are incompressible, they obey no rule – at least, no rule that f can exploit to describe them compactly. The random strings are also more numerous. Assuming that a useful encoding f will produce an f-description of a string σ that is at least $j \geq 1$ shorter than σ, very few strings usefully compress: a proportion of at most 2^{-j} strings of a given length. Even with the most pitiful amount of compression, $j = 1$, we see that at most half the strings of a given length can be compressed by any algorithm f; and the compressible (non-random) strings are sparser the more compression we demand.

An encoding f is **at least as good** as g iff there is some constant k_g such that for any string σ, $C_f(\sigma) \leq C_g(\sigma) + k_g$, so that the f-complexity of any string is within k_g of the g-complexity (where k_g is independent of σ). Kolmorogorov showed that there is an encoding algorithm f which is at least as good, in this sense, as any other algorithm (Kolmogorov 1965; Chaitin 1966; Martin-Löf 1969b). Kolmogorov called such an algorithm **asymptotically optimal**, because, since k_g is fixed independently of σ, it becomes asymptotically negligible as $|\sigma|$ increases.

Fixing on an asymptotically optimal function u, we define **complexity** simpliciter: $C(\sigma) = C_u(\sigma)$. Since u is optimal, it is at least as good as the identity function; it follows

that there exists a constant k such that $C(\sigma) \leq |\sigma| + k$. On the other hand, we also know that the proportion of strings of a given length for which $C(\sigma) \leq |\sigma| - k$ is at most 2^{-k}. As the length of the strings increases, k remains constant; so there is a length such that for all strings of at least that length, $C(\sigma) \approx |\sigma| \pm k$. Even if k is quite large, it is fixed, and for any fixed k, almost all strings are longer than k. So almost all strings are not noticeably compressible when compared to their initial length, and almost all strings have complexity of approximately their length. The *typical* finite string is incompressible. Implicitly following Laplace, Kolmogorov proposes that incompressible strings, being both unruly and typical, are the random ones: a string σ is **C-random** iff $C(\sigma) \approx |\sigma|$.

As we have placed no constraints on which descriptions are permissible descriptions, a carefully designed compression algorithm could encode more information than is in the content of the description itself: an efficient decoding

> might begin its operation by scanning all of δ to determine its length, only then to read the contents of δ bit for bit. In this way, the information δ is really worth $|\delta| + \log|\delta|$ bits, so it is clear that we have been cheating in calling $|\delta|$ the complexity of σ.
>
> (van Lambalgen 1987: p. 736; notation altered)

To exclude this possibility, we follow Chaitin and Levin and restrict the permissible descriptions to those which are **prefix-free** with respect to u (Chaitin 1966; Levin 1976). An algorithm f is prefix-free iff no two f-descriptions are such that one is an initial segment of the other. (Think telephone numbers:[5] no well-formed phone number is such that appending further digits to it yields another well-formed phone number.) With a prefix-free encoding, a decompression algorithm can begin processing the description while reading it, and recognise the end of the description without having to be explicitly told that it has ended. Accordingly, such algorithms make use of only $|\delta|$ bits in decoding a string. We can define the **prefix-free f-complexity** of a string as the length of the shortest prefix-free f-description that generates the string, and then define the **prefix-free complexity** of a sequence $K(\sigma) = K_{u^*}(\sigma)$ for some fixed prefix-free asymptotically optimal u^* (which do exist: see Downey and Hirschfeldt, 2010: ch. 3).

The requirement that descriptions be prefix-free reduces the number of permissible descriptions, and thus in general increases the complexity of sequences: $C(\sigma) \leq K(\sigma)$ (the relation between C and K is discussed in Downey and Hirschfeldt 2010: ch. 4). But then the prefix-free incompressible sequences are even more numerous than the C-random sequences, and so play the functional role of random sequences even better. Accordingly, call a string **Kolmogorov random** iff $K(\sigma) \gtrapprox |\sigma|$.

This inefficiency in K is actually of benefit, for it allows us to extend the definition of Kolmogorov randomness to infinite sequences. We can define an **infinite sequence** x as **prefix-free Kolmogorov random** iff every finite initial subsequence $x \upharpoonright i$ is prefix-free Kolmogorov random. (This definition fails for the original measure of complexity C, as C is subject to the phenomenon of **complexity oscillation**, where any sufficiently long string has non-C-random initial segments: Li and Vitanyi 2008: §2.5.)

[5] Or sentences of properly bracketed formal logic (Shapiro 2013: §2, Theorem 5).

So defined, the class of infinite Kolmogorov random sequences can be shown to be non-empty. In fact, a striking result due to Schnorr shows that it is a familiar class: **the set of infinite Kolmogorov random sequences is precisely the set of Martin-Löf random sequences** (Schnorr 1971).

This result shows a happy convergence between our two discussions of randomness. It is some evidence that we have captured the intuitive notion of randomness in a precise formal notion. Some claim Schnorr's result is good evidence for a 'Martin-Löf-Chaitin thesis' (Delahaye 1993) that identifies the intuitive notion of randomness (for infinite sequences at least) with the precise notion of **Kolmogorov/Martin-Löf (KML) randomness** – analogous to Church's thesis identifying the intuitive notion of effective procedure with Turing machine/recursive function.

In the case of Church's thesis, every precise account that is even close to intuitively plausible turns out to be equivalent (Turing machines, recursive functions, abacus machines, …). This is not true of randomness, where there are many distinct extensionally inequivalent yet precise accounts of randomness that more or less agree with intuition: from those relatively closely associated with KML randomness (such as Schnorr randomness and von Mises randomness), to other proposals of more distant pedigree, such as epistemic accounts (Eagle 2005) or accounts based on indeterminism (Hellman 1978). The dispute over whether there is a single precise notion of randomness that answers perfectly to our intuitions about random sequences can be largely sidestepped for present purposes (but see Porter 2012). KML randomness is a reasonable and representative exemplar of the algorithmic approach to randomness, and it is adopted here as a useful working definition of randomness for sequences. None of the difficulties we're about to discuss concerning the connection between randomness and chance turn on the details of which particular sequences get counted as random – most arise from the mismatch between intuitions about chance, and features common to *all* accounts of product randomness.

21.3 Chance and Randomness

The question before us now is this: what role can product randomness play in the theory of chance? Does chance require the existence – or potential existence – of random sequences, as orthodox frequentism supposes? Does the existence of a random sequence entail the existence of chance? I address both questions below.

21.3.1 Chance without Randomness

Random sequences are, if the foregoing is correct, the kinds of outputs that are typical of a certain sort of chance process: **independent, identically distributed** (i.i.d.) trials of a process with two outcomes (a so-called **Bernoulli process**). But many chance processes don't have these features, and their characteristic outputs are not product random.

Frequentism gives a central role to randomness. Probabilities only exist in those mass phenomena that can be idealized to a random sequence of outcomes, namely collectives: 'we shall not speak of probability until a collective has been defined' (von Mises 1957: p. 18). But this approach prioritizes the overall sequence over the individual outcomes; and interesting

chance phenomena that arise at the level of individual trials of the repeated processes – **single-case chances** – are overlooked (Hájek 2009; Jeffrey 1992). The thrust of the literature surrounding frequentism has been that single-case chance must be part of any satisfactory account of physical probability. The upshot for randomness is that, unless the single-case chances satisfy some significant constraints, the resulting outcome sequences need not be, and typically will not be, random.

Consider physically realistic chance processes in which the outcomes are not **independent**. One might consider the sequence of states of weather on successive days: the chance of a sunny day, given a sunny day yesterday, is higher than the unconditional chance of a sunny day. Sequences of non-independent outcomes are susceptible to gambling systems (e.g., bet on a sunny day tomorrow when it's been sunny today), and are thus non-random. Such a sequence cannot be idealized to a random collective, and so is not a suitable example of the mass phenomena that von Mises takes as the proper subject matter of the theory of probability. If we are to understand the chances involved in weather patterns, they cannot be extracted from sequences of weather outcomes. One might reasonably ask, whence do these chances derive? For there is no necessity that, whenever there is a sequence of dependent trials, there must also be some sequence of independent trials; but that is exactly what von Mises' approach demands.[6]

This illustrates a major problem facing any theory of chance which, like frequentism, makes the existence of random sequences partly constitutive of chance: namely, that many chance processes don't give rise to random sequences in the sense developed in Section 21.2. For another example, consider **biased** processes. These were not a problem for von Mises' original theory of randomness, because invariance of frequencies under place selections is independent of what the frequency is. But if we accept the conclusions of Section 21.2, that von Mises' theory is intuitively inadequate, and that a typicality approach is more successful, then we face the problem that such approaches need a probability measure with respect to which sequences are typical. Perhaps it is acceptable to appeal to the Lebesgue measure in a purely mathematical account of randomness, as it is an *a priori* symmetry of the space of sequences. But biased chances are not *a priori* features of the space of outcomes, but contingent and *a posteriori* features of a particular chance device. If randomness is partly constitutive of chance, *most* chance devices – any other than fair coins – don't usually produce random outcomes, because they usually produce outcomes sequences that are atypical with respect to the Lebesgue measure. Given that, it is extremely hard to see how such chance devices end up involving chances at all, except by offering some reduction of unequal apparent chances to some combination of basic outcomes which are equally likely. Ultimately, then, such views will end up being a version of the **classical theory of probability**,[7] because such views require that *a priori* symmetries in the outcome space underlie all random sequences, and thus all genuine chances.

[6] Actual weather supervenes on some underlying mosaic of more localized outcomes, and perhaps in the actual case the frequentist *can* find some sequence which can be idealized to a random collective. But it is surely possible that there be a situation in which there is a pattern in the underlying mosaic that suggests probabilistic dependence between those outcomes. The frequentist, shackled with attempting to fit all probabilistic phenomena into the Procrustean bed of collectives, cannot handle such a case, while more liberal theories may accommodate it – even other Humean theories which ground probability ultimately in patterns of outcomes (Lewis 1994; Loewer 2004).

[7] 'The theory of chance consists in reducing all the outcomes of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their

Because frequentism gives randomness such a significant role in the foundations of chance, many cases with which frequentism has difficulty are also cases in which there is chance without randomness. Consider Hájek's **problem of ordering**: 'Limiting relative frequency depends on the order of trials, whereas probability does not' (Hájek 2009: p. 219; see also Hellman 1978). The same is true of randomness: whether a sequence is random depends on the order of the outcomes, and randomness is not necessarily preserved under permutation of outcome order. However, whether the outcomes are the product of a chancy or random process is invariant under permutation; so here we have cases in which there is process randomness without product randomness.

The preceding examples show a systematic disconnection between certain sorts of chance process and the existence of random outcome sequences. But, given the existence of single-case chances, cases of chance without randomness arise even in the most favourable circumstances for the idea that chance requires randomness. For any particular chance process could yield an **unrepresentative** outcome sequence, which doesn't reflect the underlying chances, and in which frequencies are not good evidence for the values of the chances. Those sequences will not be typical with respect to the underlying chances, and hence non-random. Such cases show don't show any systematic failure – they are, by construction, atypical of their underlying chances. But everyone except the frequentist accepts that atypical outcomes can happen.

21.3.2 Randomness without Chance

Consider short sequences: they are not compressible, since in general a prefix-free encoding of a short sequence will be about as long as the sequence itself. Accordingly, all short sequences are (KML) random. But not all short sequences involve a random process. Since some short sequences are essentially short, consisting of unrepeatable or seldom repeatable outcomes, there are some random sequences that are essentially random despite not being chancy. So the involvement of genuine chances is not a necessary condition for randomness. Perhaps we might respond by denying that all short sequences are random, treating them as a **degenerate case** of incompressibility and restricting randomness to non-degenerately incompressible sequences. But then all short sequences turn out to be non-random; and if there is an unrepeatable chance event, there will be chance without randomness. If we consider the outcomes alone, either all short sequences are random or none of them are. But as some unrepeatable outcomes are chancy, and some are not, whichever way we opt to go with respect to randomness of the singleton sequences of such outcomes we will discover cases in which we have random processes without random products, or vice versa.

Other cases of randomness without chance arise from some of the more exotic possibilities of classical physics. One sort of possibility arises from **classical indeterminism**. It is well known that classical Newtonian particle mechanics is indeterministic: that specifying the positions and momenta of all particles at a time, together with the laws, does not suffice to determine the future evolution of the system (Earman 1986; Norton 2008). But this classical indeterminism does not involve chance; nothing in the physics *requires* a probability distribution over these undetermined outcomes, rather than simply holding

existence, and in determining the number of cases favorable to the outcome whose probability is sought.' (Laplace 1951: pp. 6–7). See also Zabell, this volume (15).

that such outcomes are nomologically possible. We could revise classical mechanics by adding a measure over outcomes (though which one?). But ordinary *unrevised* classical mechanics isn't incomplete just because it does not provide chances for our credences to reflect.

This phenomenon may yield cases of product randomness without chance. By preparing a system repeatedly in a given state that does not determine its future state, and classifying the possible outcomes into two classes, '0' and '1', it is physically possible that we obtain a random binary sequence as the system evolves over time. But we then have randomness without a chance distribution over the outcomes. The possibility of randomness requires only two distinguishable possible outcomes and the ability to produce arbitrary sequences of such outcomes. Chance requires *in addition* that some measure be ascribable to each outcome. Chance is a sort of quantitative physical possibility; since a physical theory can be indeterministic without having a probability distribution over the different possible outcomes, we can produce a random sequence by an indeterministic but non-chancy process. These sorts of case may induce us to query our initial regimentation in Section 21.1, which identified process randomness with chanciness.[8] But splitting chance and process randomness further undermines any putative connection between product randomness and chance.

Another interesting possibility permitted by classical physics is **chaotic dynamics**. A (discrete) dynamical system can be characterized by four parameters: a set of basic states X, a σ-algebra Σ on X (intuitively, the possible outcomes in the system, corresponding to regions of the basic space X), some measure μ such that $\mu(X) = 1$, and a deterministic evolution map T from X onto X, which captures the lawlike evolution of states over one time step (Berkovitz and Frigg 2006). Following Werndl (2009), a dynamical system is chaotic iff it is **mixing**: for all $A, B \in \Sigma$ (where 'T^n' denotes applying $T n$ times):

$$\lim_{n \to -\infty} \mu(T^n(B) \cap A) = \mu(B)\mu(A).$$

Suppose B is the region in which our system was found n time steps ago; a system is mixing if, in the limit as n increases, the system's having been in B becomes independent of its now being in A (Berkovitz and Frigg 2006: p. 676). In the limit, such systems are maximally unpredictable (in the sense of Eagle 2005: p. 775), because states of the distant past are increasingly credentially irrelevant for future states, assuming one sets prior credences in accordance with the μ-measure of the system (though that is consistent with the system's being quite predictable in the short run). A system is **Bernoulli** when the states of the system are independent of all past (and future) states:

$$\mu\left(T^{-1}(B) \cap A\right) = \mu(A)\,\mu(B).$$

The trajectory through the state space of a Bernoulli system is just like a sequence of fair coin tosses: the past states of the system are independent of future states, and credence in them is equal to their μ-measure. Obviously all Bernoulli systems are mixing.

What's of interest for us is that chaotic systems do not require indeterminism: the evolution map T is a one-one function, and so the time evolution is deterministic. Given

[8] That is, we may consider these sorts of case as involving process randomness because of the presence of indeterminism, even if they do not involve chance.

the additional but widely believed premise that chance requires indeterminism (Lewis 1986; Schaffer 2007: p. 120), this entails that chaos does not require chance or random processes. The motivation for saying that these systems are not process random is that the underlying classical non-probabilistic dynamics provides a physically complete account of the behaviour of any particular system – i.e., sufficient to predict, and causally explain, the behaviour of the system without need of additional supplementation.

But a Bernoulli system can generate a random sequence of outcomes, because the macroscopically-described trajectory of a chaotic system through the state space (its macroscopic history) provides resources too meagre to predict its future macroscopic behaviour. These sequences can be random: they are certainly at least Borel normal, since any complex pattern of successive states A will be an element of Σ, the limit frequency with which such patterns occur will be equal to $\mu(A)$, and the frequency with which one pattern follows any other will also be $\mu(A)$.

Sampled infrequently, even mixing systems can give rise to random sequences of outcomes. Some non-Bernoulli systems of great physical interest, such as the Lorenz model of atmospheric convection (Smith 1998: §1.4), display *rapid* mixing (Luzzatto et al. 2005). Sampled at ordinary enough intervals, therefore, such systems can provide physical plausible examples of deterministic, and arguably non-chancy, random outcomes.

In such cases, admittedly, measures that look rather like chances are involved. The measure μ is formally like a probability measure, and moreover seems to be used to regulate credences about which macrostates the system will be in – reminiscent of the way chance regulates credence (Lewis 1986). This has prompted some to *deny* the additional premise that chance requires indeterminism, and argue that measures over the state space in statistical mechanics, and other dynamical theories, should be understood as chances (Clark 1987; Loewer 2001; Meacham 2005). There are also general arguments aiming to show that chance and determinism are compatible (Eagle 2011; Glynn 2010). If these arguments are successful, these random sequences will after all be associated with a chance distribution. However, it is fair to say that not everyone is convinced that deterministic chance is possible – the received view remains that the measures involved in deterministic dynamical systems are *ignorance* measures over ensembles of macroscopically indiscriminable microstates, and so not objective chances after all. (And if there is deterministic chance, we open up the prospect of yet further examples of chance without randomness: potential cases where the deterministic laws coexist with chances, but in which sequences of outcomes generated in accordance with those laws is predictable and exploitable by a gambling system).

An example of randomness in a deterministic setting (though not that of classical physics) which does not make any appeal to a measure is given by Humphreys. Adapting the theorem of Ville's mentioned in Section 21.2, he proves that there exists 'a theory which is deterministic in the sense of Montague and which has as a model a sequence which is random in the sense of von Mises/Church' (Humphreys 1978: appendix). This isn't quite KML randomness, but it is a suggestive result. The theorem exploits the fact that the standard Montague-Earman definition of determinism requires that the evolution map of a dynamical system be a function (Earman 1986; Montague 1974); but it does not require that it be computable. Here, given the explicit construction of the evolution function, it is difficult to accept that the limit frequency in the resulting sequence can play the chance role; so we have a case, albeit rather artificial, of randomness without chance.

21.3.3 Randomness as Evidence for Chance

We've seen plausible examples in which chancy processes do not produce random sequences – either by chance, in the case of Bernoulli processes, or because the family of chancy processes includes many whose typical outcomes do not show the hallmarks of random sequences.

We've also seen examples of randomness without chance. Some of these examples, such as the statistical mechanical cases, may fail to persuade, because we needed to use objective probability measures that behave in many ways like chances; and these, it may be argued, play the chance role well enough to deserve the name. Even so, some examples of randomness without chance remain, and support the claim that chancy processes are unnecessary for random products.

Together, the examples in Sections 21.3.1–21.3.2 make the case that the distinctions drawn in Section 21.1 between product and process randomness mark genuine metaphysical distinctions, and neither is plausibly reducible to the other.

Nothing in my discussion undermines the sensible position that randomness of some sequences of outcomes is **good evidence** for the involvement of a chancy process. The appearance of a series of outcomes typical of a sequence of independent and identically distributed trials is generally good evidence that there was such a sequence of trials. Conversely, seeing a sequence of outcomes governed by a recognizable law will disconfirm any probabilistic hypothesis – not conclusively, but substantially. Since the typical outcomes of repeated equiprobable i.i.d. trials are random sequences, we ought sensibly to *expect* such processes to produce a random sequence of outcomes. Accordingly, Hellman says:

> The link, then, between mathematical and physical randomness is epistemic and only that. Observations of mathematically non-random sequences can be used to decide when further explanation in terms of as yet undiscovered causal factors is wanting. But, in no sense is any notion of mathematical randomness serving as an explication for 'ultimate physical randomness', whatever that might be.
>
> (Hellman 1978: p. 86)

Taking 'mathematical' and 'physical' to mean product and process randomness respectively, this conclusion seems inescapable.

Predictably, this epistemic connection between randomness and chance resembles that between frequencies and chance. Relative frequencies are good evidence for the chances; known chances should lead us to expect certain frequencies; and extreme frequencies might prompt us to look for non-chance hypotheses. But though frequency evidence is good evidence for chance hypotheses – and though successful accounts of chance must explain *why* it's such good evidence – frequentism is implausible as a reductive account of chance, and any proposed reduction of chance to randomness is equally implausible.

By definition, the Kolmogorov complexity function K cannot be a recursive function. Hence there is no effective positive conclusive test for randomness even of finite sequences. Even if our evidence is in fact a random sequence of outcomes, the proposition that the sequence is random will not generally be in our evidence, as (for all we know) there may be a short description of the sequence that we are unaware of. That an evidence sequence is random is no less a theoretical hypothesis about it than the hypothesis that it was produced by a chance process. In that light, it seems sensible to neglect randomness, and focus

directly on how likely a chance process is to have been involved in the production of a given sequence, conditional on the contents of that evidence sequence.[9]

That said, more sophisticated descendants of frequentism, such as Lewis' **best systems analysis** of chance (Lewis 1994; Schwarz, this volume (20)), might enable us to make more precise claims about the bearing of randomness on chance. One application of randomness in a Humean metaphysics of chance might be in analysing Lewis' notion of **fit:** namely, the proposal that a probability function fits some outcomes just when those outcomes are random with respect to that function (Section 21.2.2; see also Elga 2004). Another application might be in understanding when to invoke chances, given an austere fundamental ontology of only the 'spatiotemporal arrangement of local qualities throughout all of history, past and present and future' (Lewis 1994: p. 474). One attractive suggestion might be that we ought to adopt a probabilistic description as best balancing strength and simplicity when that underlying arrangement of local qualities is complex, in the precise sense of being Kolmogorov incompressible (Section 21.2.3). For complex sequences of outcomes preclude there being simple non-probabilistic descriptions that are strong enough to carry useful content about the outcome sequence. Yet even if randomness can play these important roles in a Humean metaphysics of chance, that falls short of the kind of reduction originally envisaged by von Mises.

Acknowledgments

For helpful comments on earlier versions of this chapter, I'd like to thank Al Hájek, an audience at a session of the Paris-Diderot Philosophy of Math Seminar on algorithmic randomness, particularly Chris Porter and Sean Walsh, students in my Oxford graduate seminar on philosophy of probability, Wo Schwarz and others in a session of the ANU philosophy of probability reading group, and colleagues from Adelaide and Flinders for comments on a written draft.

References

Berkovitz, J. and Frigg, R. (2006) The Ergodic Hierarchy, Randomness and Hamiltonian Chaos. *Studies in History and Philosophy of Modern Physics*. 37. pp. 661–91.

Borel, E. (1909) Les Probabilités Dénombrables Et Leurs Applications Arithmétiques. *Rendiconti Del Circolo Matematico Di Palermo*. 27. pp. 247–71.

Chaitin, G. (1966) On the Length of Programs for Computing Finite Binary Sequences. *Journal of the Association for Computing Machinery*. 13. pp. 547–69.

[9] We are generally in a better position to know that a process is chancy when we draw that conclusion on the basis of consideration of its physical characteristics, than when we draw that conclusion by performing a double abduction, first inferring that a sequence of outcomes is in fact random from the fact that it seems random, and then inferring that the process behind that random sequence is chancy.

The irony of concluding a chapter that spends so much time talking about randomness by recommending that it be ignored does not escape me.

Church, A. (1940) On the Concept of a Random Sequence. *Bulletin of the American Mathematical Society*. 46. pp. 130–5.

Clark, P. (1987) Determinism and Probability in Physics. *Proceedings of the Aristotelian Society. Supplementary Volume*. 61. pp. 185–210.

Dasgupta, A. (2011) Mathematical Foundations of Randomness. In Bandyopadhyay, P. and Forster, M. (eds.) *Handbook of the Philosophy of Science.* Volume 7: Philosophy of Statistics. pp. 641–710. Amsterdam: North-Holland.

Delahaye, J.-P. (1993) Randomness, Unpredictability and Absence of Order: The Identification by the Theory of Recursivity of the Mathematical Notion of Random Sequence. In Dubucs, J.-P. (ed.) *Philosophy of Probability*. pp. 145–67. Dordrecht: Kluwer.

Downey, R. and Hirschfeldt, D. R. (2010) *Algorithmic Randomness and Complexity*. New York, NY: Springer.

Eagle, A. (2005) Randomness Is Unpredictability. *The British Journal for the Philosophy of Science*. 56. 4. pp. 749–90.

Eagle, A. (2011) Deterministic Chance. *Noûs*. 45. pp. 269–99.

Eagle, A. (2014) Chance Versus Randomness. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Spring. [Online] Available from: http://plato.stanford.edu/archives/spr2014/entries/chance-randomness/. [Accessed 18 Sep 2015]

Earman, J. (1986) *A Primer on Determinism*. Dordrecht: Reidel.

Elga, A. (2004) Infinitesimal Chances and the Laws of Nature. *Australasian Journal of Philosophy*. 82. pp. 67–76.

Gaifman, H. and Snir, M. (1982) Probabilities Over Rich Languages, Testing and Randomness. *The Journal of Symbolic Logic*. 47. pp. 495–548.

Glynn, L. (2010) Deterministic Chance. *British Journal for the Philosophy of Science*. 61. pp. 51–80.

Hájek, A. (2009) Fifteen Arguments Against Hypothetical Frequentism. *Erkenntnis*. 70. pp. 211–35.

Hawthorne, J. and Lasonen-Aarnio, M. (2009) Knowledge and Objective Chance. In Greenough, P. and Pritchard, D. (eds) *Williamson on Knowledge*. pp. 92–108. Oxford: Oxford University Press.

Hellman, G. (1978) Randomness and Reality. In Asquith, P. D. and Hacking, I. (eds.) *Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 2. pp. 79–97. Chicago, IL: University of Chicago Press.

Humphreys, P. W. (1978) Is 'Physical Randomness' Just Indeterminism in Disguise? In Asquith, P. D. and Hacking, I. (eds) *Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 2. pp. 98–113. Chicago, IL: University of Chicago Press.

Jeffrey, R. C. (1992) Mises Redux. In *Probability and the Art of Judgement*. pp. 192–202. Cambridge: Cambridge University Press.

Kolmogorov, A. N. (1963) On Tables of Random Numbers. *Sankhyā*. 25. pp. 369–76.

Kolmogorov, A. N. (1965) Three Approaches to the Definition of the Concept 'Quantity of Information'. *Problemy Peredachi Informatsii*. 1. pp. 3–11.

Kolmogorov, A. N. and Uspensky, V. A. (1988) Algorithms and Randomness. *SIAM Theory of Probability and Applications*. 32. pp. 389–412.

Laplace, P. -S. (1951) *Philosophical Essay on Probabilities*. New York, NY: Dover.

Levin, L. A. (1976) Various Measures of Complexity for Finite Objects. *Soviet Mathematics Doklady*. 17. pp. 522–6.

Lewis, D. (1986) A Subjectivist's Guide to Objective Chance. In *Philosophical Papers*. Vol. 2. pp. 83–132. Oxford: Oxford University Press.

Lewis, D. (1994) Humean Supervenience Debugged. *Mind*. 103. pp. 473–90.
Li, M. and Vitanyi, P. M. B. (2008) *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd edition. Berlin and New York: Springer Verlag.
Lieb, E. H., Osherson, D., and Weinstein, S. (2006) Elementary Proof of a Theorem of Jean Ville. eprint arXiv:cs/0607054 [cs.CC].
Loewer, B. (2001) Determinism and Chance. *Studies in History and Philosophy of Modern Physics*. 32. pp. 609–20.
Loewer, B. (2004) David Lewis's Humean Theory of Objective Chance. *Philosophy of Science*. 71. pp. 1115–25.
Luzzatto, S., Melbourne, I., and Paccaut, F. (2005) The Lorenz Attractor Is Mixing. *Communications in Mathematical Physics*. 260. pp. 393–401.
Martin-Löf, P. (1966) The Definition of Random Sequences. *Information and Control*. 9. 6. pp. 602–19.
Martin-Löf, P. (1969a) The Literature on Von Mises' Kollektivs Revisited. *Theoria*. 35. pp. 12–37.
Martin-Löf, P. (1969b) Algorithms and Randomness. *Revue De l'Institut International De Statistique*. 37. pp. 265–72.
Meacham, C. J. G. (2005) Three Proposals Regarding a Theory of Chance. *Philosophical Perspectives*. 19 pp. 281–307.
Montague, R. (1974) Deterministic Theories. In Thomason, R. H. (ed.) *Formal Philosophy*. pp. 303–59. New Haven, CT: Yale University Press.
Nies, A. (2009) *Computability and Randomness*. Oxford Logic Guides 51. Oxford: Oxford University Press.
Norton, J. (2008) The Dome: an Unexpectedly Simple Failure of Determinism. *Philosophy of Science*. 75. pp. 786–98.
Porter, C. (2012) *Mathematical and Philosophical Perspectives on Algorithmic Randomness*. Ph.D. thesis. Notre Dame, IN: University of Notre Dame.
Rathmanner, S. and Hutter, M. (2011) A Philosophical Treatise of Universal Induction. *Entropy*. 13. pp. 1076–136.
Schaffer, J. (2007) Deterministic Chance? *The British Journal for the Philosophy of Science*. 58. pp. 113–40.
Schnorr, C. P. (1971) A Unified Approach to the Definition of Random Sequences. *Theory of Computing Systems*. 5. pp. 246–58.
Shapiro, S. (2013) Classical Logic. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Winter. [Online] Available from: plato.stanford.edu/archives/win2013/entries/logic-classical/. [Accessed 18 Sep 2015.]
Smith, P. (1998) *Explaining Chaos*. Cambridge: Cambridge University Press.
Solomonoff, R. (1964) A Formal Theory of Inductive Inference, Parts I and II. *Information and Control*. 7. pp. 1–22, 224–54.
Vadhan, S. P. (2012) *Pseudorandomness*. Foundations and Trends®in Theoretical Computer Science. 7. pp. 1–336.
van Lambalgen, M. (1987) Von Mises' Definition of Random Sequences Revisited. *Journal of Symbolic Logic*. 52. pp. 725–55.
Ville, J. (1939) *Étude Critique de la Notion Collectif*. Paris: Gauthier-Villars.
von Mises, R. (1957) *Probability, Statistics and Truth*. New York: Dover.
Wald, A. (1938) Die Widerspruchfreiheit des Kollektivbegriffs in der Wahrscheinlichkeitsrechnung. *Ergebnisse eines Mathematischen Kolloquiums*. 8. pp. 38–72.
Werndl, C. (2009) What Are the New Implications of Chaos for Unpredictability? *The British Journal for the Philosophy of Science*. 60. pp. 195–220.

CHAPTER 22

CHANCE AND DETERMINISM

ROMAN FRIGG

22.1 Introduction

Determinism and chance seem to be irreconcilable opposites: either something is chancy or it is deterministic, but not both. Yet there are processes which appear to square the circle by being chancy and deterministic at once, and the appearance is backed by well-confirmed scientific theories, such as statistical mechanics, which also seem to provide us with chances for deterministic processes. Is this possible, and if so how? In this chapter I discuss this question for probabilities as they occur in the natural sciences, setting aside metaphysical questions in connection with free will, divine intervention and determinism in history.

The first step is to come to a clear formulation of the problem. To this end we introduce the basic notions in play in some detail, beginning with determinism. Let W be the class of all physically possible worlds. The world $w \in W$ is deterministic iff for *any* world $w' \in W$ it is the case that: if w and w' are in the same state at some time t_0 then they are in the same state at *all* times t (Earman 1986: p. 13). The world w is indeterministic if it is not deterministic. This definition can be restricted to a subsystem s of w. Consider the subset $W_s \subseteq W$ of all possible worlds which contain a counterpart of s, and let s' be the counterpart of s in w'. Then s is deterministic iff for *any* world $w' \in W_s$ it is the case that if s and s' are in the same state at some time t_0 then they are in the same state at *all* times t. This makes room for partial determinism because it is in principle possible that s is deterministic while (parts of) the rest of the world are indeterministic. The systems formulation of determinism will facilitate the discussion because standard examples of deterministic processes occur in relatively small systems rather than the world as a whole.

To introduce chance we first have to define probabilities. Consider a non-empty set Ω. An algebra on Ω is a set Σ of subsets of Ω so that $\Omega \in \Sigma$, $\sigma_i \backslash \sigma_j \in \Sigma$ for all $\sigma_i, \sigma_j \in \Sigma$, and $\cup_{i=1}^{n} \sigma_i \in \Sigma$ if all $\sigma_i \in \Sigma$.[1] A probability function p is a function $\Sigma \rightarrow [0, 1]$ which assigns every member of Σ a number in the unit interval $[0, 1]$ so that $p(\Omega) = 1$ and $p(\sigma_i \cup \sigma_j) = p(\sigma_i) + p(\sigma_j)$ for all $\sigma_i, \sigma_j \in \Omega$ for which $\sigma_i \cap \sigma_j = \emptyset$. The requirements that p be in the

[1] If n is finite, then Σ is just an algebra. If it is closed under countable unions, it is a *sigma algebra*.

unit interval, assign probability 1 to Ω, and satisfy the addition rule for non-overlapping sets are known as the *axioms of probability*.[2] Provided that $p(\sigma_j) > 0$, $p(\sigma_i \mid \sigma_j) = p(\sigma_i \cap \sigma_j) / p(\sigma_j)$ is the conditional probability of σ_i on σ_j. If we throw a normal die once Ω is the set {1, 2, 3, 4, 5, 6}, and Σ contains sets such as {3} ('getting number 3'), {2, 4, 6} ('getting an even number') and {5, 6} ('getting a number larger than 4'). The usual probability function is $p(\{i\}) = 1/6$ for $i = 1, ..., 6$. The addition rule yields $p(\{2, 4, 6\}) = 1/2$ and $p(\{5, 6\}) = 1/3$. In what follows we will refer to the elements of Σ as *events*. This is a choice of convenience motivated by the fact that the sort of things to which we will attribute probabilities below are most naturally spoken of as 'events'.

An alternative yet equivalent approach formulates the axioms of probability in terms of propositions (or sentences). To every element of $\sigma \in \Sigma$ there corresponds a proposition $\pi[\sigma]$ which says that σ is the case. The second and third axioms of probability then say that $p(\pi[\Omega]) = 1$ where $\pi[\Omega]$ is a tautology and $p(\pi[\sigma_i] \vee \pi[\sigma_j]) = p(\pi[\sigma_i]) + p(\pi[\sigma_j])$ for all logically incompatible propositions $\pi[\sigma_i]$ and $\pi[\sigma_j]$, where '$\vee$' stands for 'or'.

An interpretation of probability specifies the meaning of probability statements. Interpretations fall into two groups: objective and subjective. Objective probabilities are rooted in objective features of the world. If, say, the objective probability of obtaining heads when flipping a coin is 0.5, then this is so because of facts in the world and not because of what certain agents believe about it or because of the evidence supporting such a claim. Objective probabilities contrast with subjective probabilities or *credences*. A credence is a degree of belief an agent has (or ought to have) in the occurrence of a certain event. We write *cr* to indicate that a certain probability is a credence.

The two most common kinds of objective probability are relative frequencies and chances. *Frequencies* are calculated in a sequence (finite or infinite) of events of the same kind and hence provide a statistical summary of the distribution of certain features in that sequence.[3] *Chances* are different from frequencies in that they apply to single cases in virtue of intrinsic properties of these cases. There is a 0.5 chance that the particular coin that I am going to flip now will land heads. But the fact that 10% of students at LSE get first-class marks does not warrant the claim that James has a 0.1 chance of getting a first-class mark for his next essay. Frequncies can be both the manifestation of, and evidence for, chances, but they are not themselves chances. To indicate that a certain objective probability *p* is a chance we write *ch*.

Let us introduce a few notational conventions. In what follows we often speak about events such as getting heads when flipping a coin. If we speak about an event informally I use '*e*' rather than 'σ' to keep notation intuitive, and I take '*E*' to be the outcome-specifying proposition saying that *e* obtains. It has become customary to attribute chances to propositions. I will follow this convention and write *ch*(*E*) for the chance that *e* occurs. Likewise *cr*(*E*) refers to the degree of belief of an agent that *e* occurs.

A chance function is nontrivial if there are events for which it assumes non-extremal values, i.e. values different from zero or one. There is widespread consensus that nontrivial chances are incompatible with determinism. In an often-quoted passage Popper (1982: p. 105) states that

[2] As is well known, there are different axiomatizations of probability. Nothing in what follows depends on which axioms we choose (in particular, nothing depends on whether probabilities are assumed to be finitely additive or countably additive). For a discussion of alternative axiomatizations see Lyon's chapter in this volume (Lyon 2016).

[3] For a discussion of frequentism see Hájek (1997) and La Caze's chapter in this volume (La Caze 2016).

> objective physical probabilities are incompatible with determinism; and if classical physics is deterministic, it must be incompatible with an objective interpretation of classical statistical mechanics,

and Lewis (1986: p. 118, original emphasis) exclaims that

> [t]o the question how chance can be reconciled with determinism […] my answer is: *it can't be done*.

Let us refer to this view as incompatibilism; conversely, compatibilism holds that there can be nontrivial chances in deterministic systems.

Incompatibilism is often asserted and seems to enjoy the status of an obvious truism, which is why one usually finds little argument for the position (we turn to exceptions below). Since incompatibilism undoubtedly has intuitive appeal, there is a temptation to simply leave it at that. Unfortunately things are more involved. In fact, there is a tension between incompatibilism and the fact that common sense as well as scientific theories assign probabilities to deterministic events. We assign nontrivial probabilities to the outcomes of gambling devices such as coins, roulette wheels, and dice even though we know that these devices are governed by the deterministic laws of Newtonian mechanics, and statistical mechanics attributes nontrivial probabilities to certain physical processes to occur even though the underlying mechanics is deterministic.[4]

A conflict with incompatibilism can be avoided if the probabilities occurring in deterministic theories are interpreted as credences rather than chances. On that view, the probabilities we attach to the outcomes of roulette wheels and the like codify our ignorance about the situation and do not describe the physical properties of the system itself. Outcomes are determined and there is nothing chancy about them; we are just in an epistemic situation that does not give us access to the relevant information.

This is unsatisfactory. There are fixed probabilities for certain events to occur, which are subjected to experimental test. The correctness of the probabilistic predictions of statistical mechanics has been assessed in countless laboratory experiments, and the owners of casinos make sure that their roulette wheels are unbiased to avoid loss when faced with attentive punters. These observations are difficult to square with a view that interprets these probabilities as credences. The chance of a roulette wheel stopping at slot No. 23 seems to have nothing to do with what we know about it, let alone with the existence of belief-forming creatures. The values of these probabilities seem to be determined by how things are and not by what anybody believes about them.[5]

[4] Let me add two caveats. First, in general, Newtonian mechanics need not be deterministic (Norton 2008). However, in the applications we are concerned with in this essay (gambling devices and statistical mechanics) the resulting laws are deterministic. Secondly, we now believe that Newtonian mechanics is just an approximation and the true underlying theory of the world is quantum mechanics, which is indeterministic (according to the standard interpretation). We can set aside the question of whether or not the true fundamental theory of the world is deterministic. What matters for the current discussion is the conceptual observation that probabilities are assigned to events in a deterministic context, and the issue is how to understand such probabilities – whether probabilities of real roulette wheels are of that kind is a different matter.

[5] This point is often made in the context of statistical mechanics; for instance see Redhead (1995: pp. 27–8). Lyon (2011) extends arguments of this kind to the probabilities we find in biology.

This leaves us with a dilemma: either we deny, the above points about empirical testing notwithstanding, that the probabilities in deterministic theories such as statistical mechanics are chances, or we reconsider incompatibilism. This chapter is about the second horn of the dilemma. In doing so I restrict attention to broadly Humean approaches to chance and set aside propensity interpretations.[6] These include various versions of what has become known as Humean Objective Chance (Sections 22.2 and 22.3) and the so-called method of arbitrary functions (Section 22.4). Throughout I will use the example of a coin, which can land either heads (H) or tails (T). This is for two reasons. First, the example is intuitively easy to grasp and yet sufficiently complex to make all the essential points. Secondly, the arguments developed with this example can be carried over without difficulties to more complicated cases, in particular in statistical mechanics.

22.2 Humean Objective Chance

The interpretation of probability now known as Humean Objective Chance (HOC) originates in the work of Lewis (1980, 1986, 1994),[7] and the entire approach in which HOCs occur is known as the Humean Best Systems (HBS) approach. Consider an outcome-specifying proposition E. The following definition then encapsulates the core idea of HOC:

The HOC of event e occurring at a particular time t in world w, $ch_{tw}(E)$, is a number in the interval [0, 1] such that

(1) $ch_{tw}(E)$ satisfy the axioms of probability;
(2) $ch_{tw}(E)$ supervene on the Humean Mosaic in the right way;
(3) $ch_{tw}(E)$ be the correct plug-in for X in the Principal Principle.

The first clause is a necessary condition for HOCs to be an interpretation of probability.[8] The second clause is more involved. The *Humean Mosaic* (HM) is the collection of everything that actually happens; that is, all occurrent facts everywhere and at all times. There is a question about how exactly occurrent facts ought to be characterized; the important point for now is that irreducible modalities, powers, propensities, necessary connections, and so forth are not part of HM. That is the 'Humean' in HOC.

Supervenience requires that chances be entailed by the overall pattern of events and processes in HM. HOCs supervene on HM, but unlike actual frequencies they don't supervene *simply*. The problem with simple supervenience is that it makes no room for

[6] For discussion of the compatibility of propensities with determinism see Clark (2001) and Berkovitz (2015). Alternative routes are explored in Eagle (2011), Ismael (2009), and Clark and Butterfield (1987).

[7] 'Humean Objective Chance' could be deemed bad terminology because chances by definition are objective. I use the term because it has become customary to refer to the kinds of chances introduced in this section as HOCs (witness the title of Lewis' 1980 paper!). For a general introduction to Best Systems approaches see Schwarz's chapter in this volume (Schwartz 2016).

[8] The presentation of HOC in this section follows see Frigg and Hoefer (2010). I set aside problems that are tangential to the issue of compatibilism. Among these are the temporality of chance, the justification of the Principal Principle, the alleged commitment of HOC to classical physics, and undermining. For a discussion of these see Hoefer (2007), Pettigrew (2012), Darby (2012), and Roberts (2001), respectively.

frequency tolerance. Imagine that in the history of the universe only one roulette wheel with 83 slots has ever been built and that it has been destroyed after having been spun only three times. The outcomes of these three spins were 25, 39, and 25. Actual frequentism commits us to saying that the chance for 25 is 2/3, the chance for 39 is 1/3, and all other chances are 0. This conclusion can be avoided if we see chances as supervening on facts not simply but instead in a Human Best System (HBS) way: chances are the numbers assigned to events by probability functions that are part of a *best system* (BS), where 'best' means that the system offers the best balance of *simplicity, strength* and *fit* that the events in HM allow.

Simplicity and strength are notoriously difficult to explicate.[9] It is sufficient for now to go with intuitive notions of simplicity and strength; we will return to the issue below in the context of the discussion of compatibilism. Fit is more straightforward. Every system assigns probabilities to possible courses of history. The fit of the system is measured by the probability that it assigns to the *actual* course of history. The more likely a system sees the actual course of history as, the better its fit. As an illustration, consider an HM that consists of just ten outcomes of a coin flip: *HHTHTTHHTT*. A system positing $ch(H) = ch(T) = 0.5$ has better fit than one that says $ch(H) = 0.1$ and $ch(T) = 0.9$ because $0.1^5 0.9^5 < 0.5^{10}$. This example also shows that a system has better fit when it stays close to actual frequencies, as we would intuitively expect.[10]

The motivation behind the Principal Principle (PP) is that chances are guides to action, and the PP establishes a connection between chances and the credences a rational agent should assign to certain events. In a nutshell the PP says that a rational agent who knows the chance of E should have credence in E that is equal to the chance of E as long as the agent has no inadmissible knowledge relating to E's truth. Let '*cr*' be an initial credence function.[11] In formal terms, PP is the rule that

$$cr(E \mid X\&K) = x \tag{1}$$

where X is the proposition that the chance of E takes the value x at time t in world w (i.e. X says '$ch_{tw}(E) = x$') and K is the agent's total evidence pertaining to E, which must not contain inadmissible elements.

The power of the PP depends on what qualifies as admissible evidence.[12] Intuitively, a proposition is inadmissible if it 'bypasses' X and provides information about E other than the information already contained in X. The most obvious case of such a proposition is E itself. If, for some reason (maybe because you have a reliable crystal ball), you know that E is true, then knowing the truth of E trumps any chance law about E and a rational agent's credence in E should be 1 no matter what the chance of E. Lewis (1986: p. 92) does not provide a definition of admissibility, but he offers a characterization: 'Admissible

[9] See Lewis (1994: pp. 478–80) for a forceful reply to this problem.

[10] This explication of fit has certain limitations. It is readily applicable only to HM's with a finite number of discrete chance events. Since events we consider in what follows are of this kind this limitation need not concern us. A generalization of the above definition to infinite sequences has been suggested by Elga (2004).

[11] See Lewis (1986: pp. 87–9). The credence function is 'initial' in the sense that it is the credence function of a hypothetical actor who is able to evaluate conditional probabilities but otherwise knows absolutely nothing – in Alan Hájek's words, it is the credence function of 'super-baby'.

[12] The correct formulation of the PP and the characterization of admissibility are fraught with controversy. I follow Hoefer (2007) in preferring Lewis' original formulation (given here). For a compact account of the various moves in the debate see Vranas (2004).

propositions are the sort of information whose impact on credence about outcomes comes entirely by way of credence about the chances of those outcomes.' Within the class of admissible propositions two kinds of propositions are of particular significance. The first is historical information. If a proposition is entirely about past matters of fact, then it is admissible. Boolean combinations of such statements are admissible too, and so it follows that at any given time t, H_{tw}, the entire history of world w up to time t, is admissible. The second kind is statements of laws of nature. As with historical propositions, Boolean combinations of laws are admissible too. L_w, the conjunction of all laws of nature in w, is therefore admissible.

22.3 Incompatibilism Scrutinized

We will now see that one's stand on compatibilism depends on how the details of the above are fleshed out. Recall from the introduction, Lewis (1986: p. 120) was an advocate of incompatibilism, but failed to provide an argument for his position and instead merely asserted the point: 'There is no chance without chance. If our world is deterministic there are no chances in it, save chances of zero and one.' Several authors have tried to fill this gap. Hoefer (2007: pp. 558–9) and Schaffer (2007: p. 128) provide the following reconstruction of the incompatibilist's argument. Let $ch_{tw}(E) = x$ be the nontrivial chance of E in world w at time t. As we have seen in the last section, Lewis regards historical facts and laws as admissible. The PP then tells us that

$$cr(E|\, ch_{tw}(E) = x \,\&\, H_{tw} \,\&\, L_w) = x$$

Assume that E is true.[13] If w is deterministic, then H_{tw} and L_{tw} logically imply E. Hence the axioms of probability dictate that

$$cr(E|\, ch_{tw}(E) = x \,\&\, H_{tw} \,\&\, L_w) = 1$$

So under determinism credences should be 0 and 1, but this is in contradiction with the PP.

If we stick with Lewis' view that H_{tw} and L_{tw} are admissible, then the only way to avoid the contradiction is to deny that there is a nontrivial chance for E. Recently Schaffer (2007) has provided further reasons for thinking that this is the right response. He formulates six platitudes about chance and then argues that these platitudes cannot hold true in a deterministic setting. His platitudes are the Principal Principle, the Basic Chance Principle, the Lawful Magnitude Principle, the Intrinsicness Requirement, the Future Principle, and the Causal Transition Constraint. We here focus on the first four.[14]

The first of Schaffer's platitudes is the PP itself, which, as we have just seen, leads to a contradiction when applied in deterministic setting. Since he grants the PP the status of a platitude, the PP itself cannot be given up, which leads us to the conclusion that only trivial chances are compatible with determinism.

[13] This can be done without loss of generality. The argument is *mutatis mutandis* the same if 'not E' rather than E is true.

[14] We set aside the last two because, contra Schaffer, the Future Principle and the Causal Transition Constraint are not platitudes about chance. See Hoefer (2007: pp. 554–5) for a discussion of the former; Glynn (2010: pp. 75–6) argues against the latter.

The *Basic Chance Principle*, originally due to Bigelow, Collins, and Pargetter (1993), asserts that if at time t there is a nontrivial objective chance for E in world w, then there is a possible world with the same history as w up to t and in which E is true. Schaffer proposes a stronger version of this principle, his *Realization Principle*, which adds the requirement that the possible world in which E is true must have the same laws as w. If $ch_{tw}(E)$ is nontrivial, then both $ch_{tw}(E)$ and $ch_{tw}(\neg E)$ are strictly between 0 and 1 ('¬' stands for 'not'), which implies that there is at least one possible world with the same history and laws as our world in which E is true, and likewise there is at least one such possible world in which $\neg E$ is true. This, however, is precisely what determinism denies. Hence, nontrivial chances are incompatible with determinism.

The *Lawful Magnitude Principle* codifies the view that chance values should fit with the values given by the laws of nature. If there is a chance for the coin to land heads, then this chance has to follow from the laws of nature. But if the laws are deterministic, they cannot imply nontrivial probabilities ('no probability in, no probability out').

The *Intrinsicness Requirement* says that if you have physically identical setup conditions at two different times, then the chances of their corresponding possible outcomes must be the same. This platitude is violated in a deterministic world because we only have the same setup conditions if the system's initial state is the same, but by determinism same initial conditions lead to same outcomes, which rules out nontrivial chances.

This, thinks Schaffer, seals the case against compatibilism. To see how one could counter this argument it is instructive to notice in what way exactly determinism and chance are at loggerheads. The conflict is not one of *simple* inconsistency: there are no chance laws covering exactly the same events as deterministic laws. The conflict arises if we accept reductive relations. There is a chance law for coins, and there is a deterministic mechanical theory for elementary particles. These two are inconsistent once we assume that coins are made up of atoms and that the behaviour of the coin is therefore determined by the behaviour of the atoms.

The above arguments for incompatibilism are based on giving primacy to fundamental laws wherever they are in conflict with (purported) non-fundamental laws (such as chance laws for gambling devices). This reaction is closely tied to Lewis' metaphysics, which sees the world as consisting of a manifold of spacetime points which instantiate perfectly natural monadic properties. And this is all there is: '"how things are" is fully given by the fundamental, perfectly natural, properties and relations those things instantiate' (Lewis 1994: p. 474). What counts as a perfectly natural property is determined by physics, which 'is a comprehensive theory of the world, complete as well as correct. The world is as physics says it is, and there's no more to say' (Lewis 1999: pp. 33–4). Even though Lewis rarely says so explicitly, the emphasis is clearly on *fundamental* physics: the world is how fundamental physics says it is. Let us call this position *Lewisian physicalism*.

On the basis of this metaphysics, denying that there are chances if the fundamental laws of physics are deterministic is a stringent move. If E is about a perfectly natural property, then by assumption there cannot be a chance law for it. If, by contrast, E is about a property that is not perfectly natural (such as *coin* and *roulette wheel*), then a best system will not contain laws for E-type events (and a fortiori no chance laws) because a best system does not say anything about E-type events at all.

This austere elegance comes at a price: there are no laws about any non-fundamental kinds. Where we *seem* to have such laws, this is an illusion. Generalizations that look like

laws are in fact mere rules of thumb for feeble beings incapable of applying fundamental laws to complex situations.

A number of authors felt that this was too high a price to pay. A view of chance that denies the status of chance not only to probabilities attached to gambling devices, but also to the probabilities we find in macrophysics, genetics, engineering, meteorology, and many other non-fundamental sciences, has thrown out the baby with the bathwater. These probabilities codify information about the world and are subject to empirical test. This, so the thought continues, indicates that they are chances.

Those intending to pursue this line of argument have to provide a reformulation of the HBS approach, that departs from Lewis' formulation in at least two respects. First, they have to argue that non-fundamental laws are part of the best system. Secondly, they have to reformulate PP so that the above contradiction no longer arises. Different approaches differ in how they achieve these goals.

The first to present such a view was Loewer (2001, 2004). His account focuses on Bolzmannian statistical mechanics (BSM) as presented by Albert (2000): the package of Newtonian mechanics, the Past Hypothesis, and the Statistical Postulate.[15] In this approach the relevant system is considered to be the entire universe. Very roughly, the Past Hypothesis says that the universe started in a low-entropy state, which is associated with a certain small part Γ_p of the world's entire phase space. Newtonian Mechanics provides the time evolution φ_t, specifying how a point x of the universe's phase space evolves over time. The Statistical Postulate says that we should assume a uniform probability distribution over Γ_p at time t_0, the time of the Big Bang, and generate probabilities for the system's state being in any other part of the universe's phase space at any later time $t > t_0$ by conditionalizing on the past hypothesis and the system's dynamics.[16] Loewer submits that this package is a best system (2001: p. 618; 2004: p. 1124). His reasons for thinking so are that adding the Past Hypothesis and the Statistical Postulate to Newtonian Mechanics results in a powerful system:

> It is simple and it is enormously informative. In particular, it entails probabilistic versions of all the principles of thermodynamics. That it qualifies as a best system for a world like ours is very plausible.
>
> (Loewer 2001: p. 618)

The probabilities generated by the theory's Statistical Postulate therefore are chances. He calls them 'macro chances', indicating that a revision of the PP is needed (Loewer 2001: p. 618). The modified principle, 'PP(macro)', differs from the PP in that it posits that '[m]ore detailed information about the micro-condition at t than that given by the macro-condition at t is macroscopically inadmissible' (Loewer 2001: pp. 618–19). Regimenting admissible information successfully blocks the above contradiction, and hence removes the incentive to deny the existence of chances in a deterministic world (we will return to Schaffer's platitudes below).

There are a number of challenges for Loewer's approach, which are all rooted in the fact that the work in Loewer's approach is done entirely by an appeal to simplicity. The first

[15] Loewer also includes Bohmean quantum mechanics in his discussion. For want of space I discuss only statistical mechanics here, but nothing is lost since (as Loewer points out) the arguments are entirely parallel for the two cases.

[16] For an extensive discussion of statistical mechanics see Frigg (2008), Uffink (2006), and Myrvold's chapter in this volume (Myrvold 2016).

challenge is to justify why one would introduce macro chance laws at all. In BSM macrostates supervene on microstates and so a Lewisian physicalist may insist that BSM macrostates are as dispensable in the best system of the world as coins, roulette wheels, genes, steel fatigue, cloud albedo, or any other not perfectly natural property. Loewer's motivation for introducing macrostates is that describing a tremendously complex swarm of molecules as a gas having a certain macrostate simplifies things enormously and hence adding that bit of conceptual machinery to fundamental mechanics greatly simplifies the entire system. But this argument implicitly invokes a notion of simplicity that has a computational component built into it. In principle the behaviour of the gas is completely determined by the behaviour of its constituent molecules. But it would be tremendously complicated to make predictions about the gas in terms of molecules, and having the conceptual tool of macrostates at hand allows us to say much of what we actually want to say about swarms of molecules at relatively low cost.

This argument only gets off the ground if the computational cost incurred in deriving a desired result is at least part of what we mean by simplicity; Frigg and Hoefer (2015) call this 'simplicity in derivation'. This is not a notion of simplicity a Lewisian physicalist finds natuarally appealing. One would either have to show that a Lewisian physicalist should accept such a notion or else argue that the position ought to be rejected altogether. Whichever of these options one choses, the further question then is: why stop short at BSM? The probabilistic laws of genetics seem to bring equal computational simplifications with them as the laws of BSM, so why not include these in the best system too? And so on for many laws in the special sciences.

A related worry is that postulating a uniform distribution over Γ_p for an event that happens only once in the entire history of the universe, namely the Big Bang, seems conceptually problematic even if one takes frequency tolerance seriously. It is difficult to see how such a distribution could be seen as supervening on HM, and the only reason to accept it at all is its simplifying power. This is an uncomfortable move as long as we operate with an unexplained notion of simplicity.

Even if we were prepared to set these worries aside, there is another problem lurking. One can show that the fit of a system can be improved by choosing a distribution that is peaked over the actual initial condition (Frigg 2008, 2010). A peaked distribution is not less simple than a uniform one (or at any rate only marginally less simple) while it greatly increases the fit of the system, and so the best system would contain a peaked rather than a uniform distribution. Countering this objection would involve arguing that peaked distributions come at simplicity costs that outweigh all gains in fit, which tells against their inclusion in a best system. But it is hard to see how such an argument could be made as long as no tight notion of simplicity is in place.

Alternative compatibilist accounts have recently been proposed by Glynn (2010) and Frigg and Hoefer (2010, 2015). These accounts have in common that they regard chances as situated at particular levels (for instance the level of genes), and they endorse a thoroughgoing pluralism which allows for chances to occur (at least in principle) at every level (and not only at the level of BSM). They differ in how they justify and develop their position. Glynn (2010: p. 57) posits that 'there exist probabilistic high-level or special scientific laws even in such worlds [i.e. fundamentally deterministic worlds]' and points out that '[t]he probabilities projected by these laws should be regarded as genuine, objective chances because the laws in question are genuine, objective laws'. His reason for regarding special science laws as genuine laws are similar to Fodor's (1974) reasons to support the autonomy of the special sciences: higher-level properties are typically multiply realizable

at the macro level, and therefore laws about higher-level kinds cannot follow from micro properties alone. This is because the micro theory by itself does not tell us which micro kinds fall into the class of realizers for a certain macro property, and hence is unable to ground generalizations about macro kinds. Therefore, adding special science laws to a system of laws makes that system more informative and stronger, and the probabilities in these laws are chances in Lewis' sense (Glynn 2010: pp. 58–9). The above contradiction (between determinism and nontrivial chance) is avoided by stipulating that a complete conjunction of all laws together with information about facts at more fundamental levels is inadmissible (Glynn 2010: pp. 68–70).

Frigg and Hoefer (2015) extend this line of argument in two ways. First they provide a more extensive argument for the conclusion that non-fundamental laws are part of a best system. The first step of their argument draws a parallel between the issue of compatibilism and the philosophy of mind. They then argue that Lewisian physicalism is untenable for reasons similar to those put forward against eliminativism. Rejecting eliminativism makes room, at least in principle, for there to be laws formulated in non-fundamental terms. To show that at least some laws of that kind are also part of the best system, they distinguish between numerical simplicity (measured in terms of the number of different laws a system contains), simplicity in derivation (roughly, the computational costs incurred in deriving a desired result), and simplicity of formulation (roughly, the ease with which a law can be formulated). They argue that the gain in simplicity in derivation due to the introduction of non-fundamental laws far outweighs the costs in numerical simplicity, while the scope and the fit of the system remain constant. For this reason such laws should be included in a best system.

The second amendment is a requirement of coherence: whenever two parts of the HBS have the same (or sufficiently overlapping) domains of application, then there must be a Humean account of how their prescriptions relate to one other. In the case where a chance law covers events that are also covered by deterministic laws, the less fundamental law must supervene in a Humean best systems way on the facts in the domain of the more fundamental law. The requirement is best illustrated with the above example of SM. Frigg and Hoefer take laboratory systems as the subject matter (rather than the entire universe). Throughout HM there are many copies of every system and so one can look at the distribution of initial conditions over Γ_p of these systems. The postulate that the best system ought to contain a uniform distribution over Γ_p then has to be justified by arguing that such a distribution is in fact the best summary of the distribution of actual initial conditions. If the conditions are spread out more or less evenly, this arguably is the case. But if it is the case that all points are concentrated in one corner of Γ_p, then a uniform distribution is not the right one to choose.

This addresses the worries that arose in connection with Loewer's account. It remains to be shown that Schaffer's platitudes can be dealt with successfully. The first one has been dealt with by altering our understanding of admissibility (which, compatibilists insist, does not require changing Lewis' characterization of admissibility; rather it requires making proper use of that characterization). Consistency with the realization principle is restored by reformulating the principle so that the scope of the principle is restricted to histories *at the relevant level*: the 'same history up to *t*' refers to the *admissible* history. Intrinsicness is dealt with along the same lines: restrict sameness of the setup to sameness with regard to admissible properties. The Lawful Magnitude Principle turns out not to be a platitude at all; in fact it is a statement of physicalism in disguise. It is the whole point of the approaches

of Loewer, Glynn, and Frigg and Hoefer that not all chances are deductive consequences of fundamental laws.

Incompatibilists remain unconvinced. But rather than quibbling about the particulars of any of the above moves, they argue that the probabilities thus introduced simply aren't chances after all. In this vein Lyon (2011) argues that the chances we find in BSM (and, needless to say, theories of gambling devices) are not chances. His reasons are that he chooses to 'stick with the usage of "chance"' that Lewis, Schaffer, and others prefer. This is because once we lay out all the platitudes we seem to have about chance, it appears that indeterministic conceptions of chance satisfy these platitudes better than deterministic ones can' (Lyon 2011: p. 420). Lyon (2011: p. 429) is quick to point out that this does not turn BSM probabilities into credences and argues that probabilities like these are of a third kind which is neither chance nor credence and which he calls *counterfactual probability*.

We have reached an impasse. If one thinks that 'chance' *means* something like propensity, primitive fundamentally chancy laws of nature, or other kinds of fundamental modalities, then one will follow Lewis and dismiss the compatibilist's probabilities as a 'kind of counterfeit chance', which is 'quite unlike genuine chance' (Lewis 1986: p. 120). If, on the other hand, one is not committed to such a view, then compatibilism is a live option and the above proposals deserve consideration. There is no ultimate right and wrong in the use of a word, and depending on one's other philosophical commitments one *can* go either way. If nothing else, the above discussion has at least shown where compatibilists and incompatibilists part ways.

22.4 The Method of Arbitrary Functions

Let us now briefly turn to an approach that is broadly Humean (in that it does not appeal to propensities, powers, and the like) but does not stand in the Lewisian best systems tradition: the method of arbitrary functions (MAF). The method was introduced by Poincaré in 1896 and has subsequently been developed by a number of eminent mathematicians, among them Borel, Fréchet, Hopf, and Khinchin.[17] Recently Strevens (2011) and Myrvold (2011) have, in different ways, appealed to the method to make sense of objective probabilities in physics.

MAF is a mathematical technique to determine a unique probability distribution for the outcomes of mechanical games of chance, or the evolution of deterministic mechanical systems more generally. A discussion of the entire theory is beyond the scope of this chapter; we restrict ourselves to illustrating the main ideas with the example of the coin. The mechanical state of a coin can be described by two sets of variables: the angle at which the coin stands with respect to the ground and the angular velocity ω (how fast the coin rotates), and its height above ground and the vertical velocity v (how fast it is thrown upwards when tossed). Assuming the coin is tossed vertically and gravity is the only force acting on it, one can classify the initial conditions according to the outcomes they will produce. If the coin is always tossed at the same height and with the same initial angle, variations in ω and v alone determine the outcome completely (because the movement of a coin is deterministic). Keller (1986) has done the calculations and the result is the graph shown in Figure 22.1, where black initial conditions result in tails and white ones in heads.[18]

[17] Von Plato (1983) provides a historical introduction to the method.

[18] The graph is a reproduction from Diaconis (1998: p. 803).

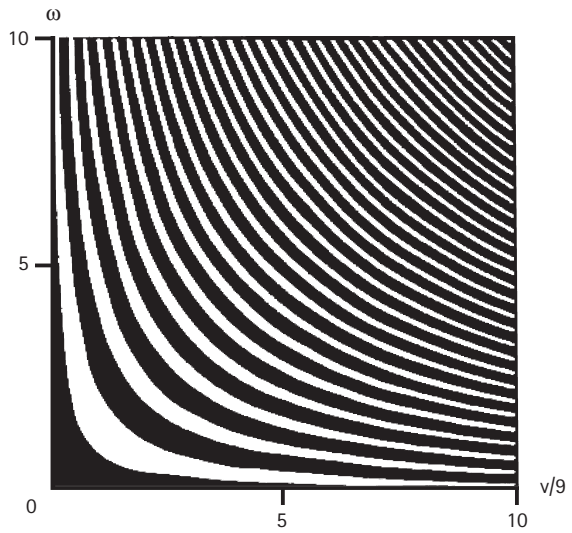


FIGURE 22.1 Black initial conditions result in tails; white ones in heads.

Assuming a probability distribution $\rho(v, \omega)$ over initial conditions, the probability for heads is just the probability of the initial conditions resulting in heads, which can be calculated by integrating $\rho(v, \omega)$ over the white regions (and *mutatis mutandis* for tails). MAF comes into play when we ask how the result of these calculations depends on our choice of $\rho(v, \omega)$. The basic result MAF seeks to establish is that for 'reasonable' $\rho(v, \omega)$ the result does not depend on the precise shape of $\rho(v, \omega)$; that is, the end result is the same for all $\rho(v, \omega)$ unless we start with a completely unreasonable $\rho(v, \omega)$ (we expect to retrieve the usual rule saying that the chance for either heads or tails is 0.5). The crucial question is of course what counts as reasonable, and the answer depends (as one would expect) on the particulars of the situation. Poincaré's original example was a roulette wheel and he argued that we obtain the usual probabilities as long as the initial probability distribution does not fluctuate wildly (in technical terms: the modulo of the derivative of the distribution has to be bounded). This approach will work here too. As we see in Figure 22.1, the black and white lines are relatively fine, and they become finer as we move towards higher ω and v; unless $\rho(v, \omega)$ fluctuates strongly on a scale of the width of the stripes, the result of the integration will not depend much on the concrete shape of $\rho(v, \omega)$. An arbitrary $\rho(v, \omega)$ will do to determine the probabilities for heads or tails – this insight gave the method its name.

In sum, MAF shows that an (almost) arbitrary distribution yields the same outcome probabilities under a deterministic physical dynamics, which seems to provide a reconciliation of determinism and chance. This enthusiasm is premature. MAF does not create probabilities *ex nihilo*; MAF only shows that outcome probabilities are invariant under the variation of *given* input probabilities. For this reason MAF per se does not ground a particular interpretation of probability, let alone provide us with a notion of deterministic chance. Whether or not MAF assists a reconciliation of chance and determinism depends on where probability distributions over initial conditions come from.

Savage (1973) interprets both input and output probabilities as credences and hence denies that MAF plays any role in reconciling chance and determinism. Myrvold (2011: p. 76) follows Savage in interpreting input probabilities as credences, but sees outcome

probabilities as *epistemic chances*. This choice of terminology indicates that outcome probabilities are determined both by epistemic and physical considerations, namely the initial credence function and the dynamics of the system. Since both are essential and irreducible, the resulting conception of probability is a combination of epistemic and physical factors. The chance aspect of these probabilities is highlighted by the fact that they are taken to satisfy a principle that is structurally similar to the PP.

There is a question, though, whether the chance aspect of Myrvold's epistemic chances is strong enough to ground a reconciliation of chance and determinism. A sceptic could argue that what MAF shows is that all reasonable initial credence functions converge towards the same credence function under the system's dynamics. This shows that all reasonable agents must have the same outcome probabilities, but it does not show that the outcome probabilities are anything other than credences: credences in, credences out. Outcome probabilities are physically constrained credences, but credences nevertheless.

Strevens takes a different line and aims to interpret MAF probabilities as physical probabilities without an epistemic component. He interprets the initial distribution $\rho(v, \omega)$ as expressing properties of the frequencies of initial conditions. Such frequencies are facts about physical world, but Strevens (2011: p. 350) emphasizes that these need not be *probabilistic* facts. Nothing forces us to interpret frequencies as probabilities; and given all the well-known difficulties of frequentism, interpreting frequencies as probabilities is best avoided. But this does not prevent us from regarding the outcome distribution as a probability distribution. So the crucial move in Strevens' account is to regard only the outcome of a process covered by MAF as probabilitiy. Strevens (2011: p. 351) calls these probabilities 'microconstant probabilities'. They are physical probabilities in that they are determined solely by facts about the frequency of initial conditions and properties of the system's dynamics.

Even though Strevens' account seems to come close to a reconciliation of chance and determinism, it is not clear whether we have passed the finishing line. Strevens never refers to microconstant probabilities as chances; nor does he ever discuss the relation between microconstant probabilities and chances. It is therefore unclear whether his account advances the compatibilist's cause. However, not much seems to be needed to turn microconstant probabilities into Humean chances (as defined by compatibilists). Borrowing from the best systems account the idea of coherence one could say that $\rho(v, \omega)$ should be chosen such that it provides the best summary of actual initial conditions, and MAF then shows that outcome probabilities do not depend sensitively on our standards of simplicity (which will determine which function we fit to the actual points). Once $\rho(v, \omega)$ is understood in this way, MAF probabilities can at least in principle be understood as Humean chances in the compatibilists' sense, which would justify their status as chances.

Acknowledgements

I would like to thank Alan Hájek, Chris Hitchcock, Carl Hoefer, Barry Loewer, Wayne Myrwold, Aidan Lyon, Hlynur Orri Stefánsson, and Michael Strevens for comments on earlier drafts and/or helpful discussions. Research for this chapter has been supported

by Grant FFI2012-37354 of the Spanish Ministry of Science and Innovation (MICINN). Thanks to Anna Warm for preparing the manuscript for publication.

References

Albert, D. (2000) *Time and Chance*. Cambridge, MA and London: Harvard University Press.

Berkovitz, J. (2015) The Propensity Interpretation of Probability: A Re-Evaluation. *Erkenntnis*. 80. 3. pp. 629–711.

Bigelow, J., Collins, J., and Pargetter, R. (1993) The Big Bad Bug: What Are the Humean's Chances? *British Journal for the Philosophy of Science*. 44. pp. 443–62.

Clark, P. (2001) Statistical Mechanics and the Propensity Interpretation of Probability. In Bricmont, J., Dürr, D., Galavotti, M. C., Ghirardi, G. C., Petruccione, F., and Zanghì, N. (eds.) *Chance in Physics: Foundations and Perspectives*. pp. 271–81. Berlin and New York: Springer.

Clark, P. and Butterfield, J. (1987) Determinism and Probability in Physics. *Proceedings of the Aristotelian Society, Supplementary Volumes*. 61. pp. 185–243.

Darby, G. (2012) Relational Holism and Humean Supervenience. *British Journal for the Philosophy of Science*. 63. 4. pp. 773–88.

Diaconis, P. (1998) A Place for Philosophy? The Rise of Modeling in Statistical Science. *Quarterly of Applied Mathematics*. 56. 4. pp. 797–805.

Eagle, A. (2011) Deterministic Chance. *Noûs*. 45. 2. pp. 269–99.

Earman, J. (1986) *A Primer on Determinism*. Dordrecht: Reidel.

Elga, A. (2004) Infinitesimal Chances and the Laws of Nature. *Australasian Journal of Philosophy*. 82. pp. 67–76.

Fodor, J. (1974) Special Sciences and the Disunity of Science as a Working Hypothesis. *Synthese*. 28. pp. 77–115.

Frigg, R. (2008) Chance in Boltzmannian Statistical Mechanics. *Philosophy of Science*. 75. 5. pp. 670–81.

Frigg, R. (2010) Probability in Boltzmannian Statistical Mechanics. In Ernst, G. and Hüttemann, A. (eds.) *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*. pp. 92–118. Cambridge: Cambridge University Press.

Frigg, R. and Hoefer, C. (2010) Determinism and Chance from a Humean Perspective. In Dieks, D., Gonzalez, W., Hartmann, S., Weber, M., Stadler, F., and Uebel, T. (eds.) *The Present Situation in the Philosophy of Science*. pp. 351–72. Berlin and New York: Springer.

Frigg, R. and Hoefer, C. (2015) The Best Humean System for Statistical Mechanics. *Erkenntnis*. 80. pp. 551–74.

Glynn, L. (2010) Deterministic Chance. *British Journal for the Philosophy of Science*. 61. 1. pp. 51–80.

Hájek, A. (1997) 'Mises Redux' – Redux. Fifteen Arguments against Finite Frequentism. *Erkenntnis*. 45. pp. 209–27.

Hoefer, C. (2007) The Third Way on Objctive Probability: A Sceptic's Guide to Objective Chance. *Mind*. 116. 463. pp. 549–96.

Ismael, J. T. (2009) Probability in Deterministic Physics. *Journal of Philosophy*. 106. 2. 89–108.

Keller, J. B. (1986) The Probability of Heads. *American Mathematical Monthly*. 93. 3. pp. 191–7.

La Caze, A. (2016) Frequentism. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Lewis, D. (1980) A Subjectivist's Guide to Objective Chance. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability.* Vol. 2. Berkeley: University of California Press. [Reprinted in Lewis 1986 pp. 83–132, with postscripts added.]
Lewis, D. (1986) *Philosophical Papers*. Vol. 2. Oxford: Oxford University Press.
Lewis, D. (1994) Humean Supervenience Debugged. *Mind*. 103. pp. 473–90.
Lewis, D. (1999) *Papers in Metaphysics and Epistemology*. Cambridge: Cambridge University Press.
Loewer, B. (2001) Determinism and Chance. *Studies in History and Philosophy of Modern Physics*. 32. pp. 609–29.
Loewer, B. (2004) David Lewis' Humean Theory of Objective Chance. *Philosophy of Science*. 71. pp. 1115–25.
Lyon, A. (2011) Deterministic Probability: Neither Chance nor Credence. *Synthese*. 182. pp. 413–32.
Lyon, A. (2016) Kolmogorov's Axiomatization and Its Discontents. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Myrvold, W. C. (2011) Deterministic Laws and Epistemic Chances. In Ben-Menahem, Y. and Hemmo, M. (eds.) *Probability in Physics*. pp. 73–85. Berlin: Springer.
Myrvold, W. C. (2016) Probabilities in Statistical Mechanics. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy.* Oxford: Oxford University Press.
Norton, J. (2008) The Dome: An Unexpectedly Simple Failure of Determinism. *Philosophy of Science.* 75. 5. pp. 786–98.
Pettigrew, R. (2012) Accuracy, Chance, and the Principal Principle. *Philosophical Review*. 121. 2. pp. 241–75.
Popper, K. R. (1982) *The Quantum Theory and the Schism in Physics*. London: Hutchinson.
Redhead, M. (1995) *From Physics to Metaphysics*. Cambridge: Cambridge University Press.
Roberts, J. T. (2001) Undermining Undermined: Why Humean Supervenience Never Needed to be Debugged (Even if it's a Necessary Truth). *Philosophy of Science*. 68. 3. (Supplement). pp. S98–S108.
Savage, L. J. (1973) Probability in Science: A Personalistic Account. In Suppes, P. (ed.) *Logic, Methodology and Philosophy of Science.* IV. pp. 417–28. Amsterdam: North-Holland.
Schaffer, J. (2007) Deterministic Chance? *British Journal for the Philosophy of Science*. 58. 2. pp. 113–40.
Schwarz, W. (2016) Best System Approaches to Chance. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Strevens, M. (2011) Probability out of Determinism. In Beisbart, C. and Hartmann, S. (eds.) *Probability in Physics*. pp. 339–64. Oxford: Oxford University Press.
Uffink, J. (2006) Compendium of the Foundations of Classical Statistical Physics. In Butterfield, J. and Earman, J. (eds.) *Philosophy of Physics*. pp. 923–1047. Amsterdam: North Holland.
von Plato, J. (1983) The Method of Arbitrary Functions. *British Journal for the Philosophy of Science*. 34. pp. 37–47.
Vranas, P. B. M. (2004) Have Your Cake and Eat It Too: The Old Principle Reconciled with the New. *Philosophy and Phenomenological Research*. 69. 2. pp. 368–82.

PART V

PROBABILISTIC JUDGMENT AND ITS APPLICATIONS

CHAPTER 23

HUMAN UNDERSTANDINGS OF PROBABILITY

MICHAEL SMITHSON

23.1 Probabilities as Degrees of Belief

The concept of a probability as reflecting a degree of belief is the principal connection between probability theories and cognitive psychology. It is all too easy to forget that the concept of probability is historically and culturally specific, and its connection with psychological uncertainty even more so. There still exist cultures with no identifiable notion of probability, and some people in Western cultures disavow any connection between probability and mental aspects of uncertainty such as degrees of belief. This chapter therefore begins with a survey of the development of this connection.

The earliest scholar to link probability with degree of belief is most likely Jacob Bernoulli in his 1713 *Ars Conjectandi*, with Laplace, De Morgan, and Donkin elaborating this view during the first half of the 19th century. Keynes (1921) is noted for having added that subjective probability judgments are logical relations between one set of propositions and another, conditioned in some sense by the knowledge available to the judge. In making this connection, Keynes could account for two reasonable judges assigning different probabilities to the same prospect by differing in the information available to them. Keynes also famously declared that not all logical probabilities are quantifiable or even comparable with one another. As Ramsey (1926/1931) observed, this claim created difficulties for the one-to-one correspondence between degrees of belief and degrees of probability relations.

Borel (1924/1964) and Ramsey disputed Keynes on this latter point and related matters. Ramsey began by arguing that there is no apparent distinction between so-called "quantifiable" and "unquantifiable" beliefs, although he allowed that some beliefs can be measured more accurately and/or reliably than others. Importantly, he dismissed introspection as a valid and accurate source of such measurements and turned, instead, to what is now the standard behavioral approach based on betting-rates, generalizing it to an account based on preferences. He thereby claimed to have found a "purely psychological" way of measuring degrees of belief.

It is at this point that Ramsey raised the question of what constitutes "reasonable" degrees of belief. From then on, concepts of coherence were elaborated, for example, by Ramsey and de Finetti (1937/1964), either in the sense of avoiding sure loss or a stronger sense than that. Subjective probability theories became more prescriptive and less descriptive. Neoclassical economics sprang directly from these developments.

A relatively independent contemporaneous contribution came from the invention of game theory, again initiated by Borel (1921/1953) and then famously elaborated by von Neumann and Morgenstern (1944). In his 1924 article, Borel made this striking prediction:

> The deep study of certain games will perhaps lead to a new chapter of the theory of probabilities... It will be a new science where psychology will be no less useful than mathematics, but this new chapter will be added to previous theories without modifying them.
>
> (Borel, E. 1924/1964: pp. 58–9)

Nevertheless, it was not until the 1950s that psychologists (principally Ward Edwards, 1954 and 1961, and John Cohen 1954 and 1960) saw a need for a non-prescriptive, psychologically informed, understanding of human probability judgments.

The nature of the boundary between the prescriptive and descriptive turned out to be contestable. Cohen (1960: pp. 28–9) wrote to de Finetti expressing the view that the psychological study of probability does not involve attributions of "error" to humans. Cohen quoted de Finetti's response, which begins "Unlike almost all mathematicians, I agree completely with your statement that every probability evaluation is a probability evaluation, that is, something to which it is meaningless to apply such attributes as right, wrong, rational, etc." Immediately thereafter, however, de Finetti makes it clear that he regards incoherent bets as "nonsensical", by which he appeared to mean "irrational".

Nonetheless, it still seemed possible that human probability judgments might correspond to the laws of probability in some respects, perhaps along the lines of the Weber-Fechner laws of human perceptions of physical properties. Indeed, a tradition of modeling subjective probabilities via continuous weighting functions survives to this day. However, studies by Cohen and his colleagues in the 1950s and 1960s (summarized in Cohen 1964) of probability estimation and reasoning in children and adults suggested that even adult judgments not only are miscalibrated but also deviate substantially from probability theory. Several of their findings anticipated later research, such as overestimation of low-probability events, intransitive choices among gambles, and the conjunction fallacy (assignment of higher probability to the conjunction of two events than to either constituent event).

A decade later, Amos Tversky and Daniel Kahneman began publishing research that became known as the "heuristics and biases" school, highlighting what they portrayed as human cognitive illusions and errors in probability judgments (e.g., Tversky and Kahneman 1974, Kahneman, Slovic, and Tversky 1982). The literature on this topic is large, and so is the list of "errors" (Hogarth 1980 reviewed 27). We shall examine several of them in sections to come. Some researchers underpinned claims that heuristics and biases are genuinely irrational by identifying correlations between scores on tests of mental abilities and the tendency to use normative strategies in judgment and decision-making (Stanovich and West 2000). Cohen, Edwards, and Herbert Simon disagreed (for somewhat different reasons) with the heuristics and biases school's emphasis on human errors. During the 1980s and

1990s a variety of debates ensued about whether particular human deviations from standard probability theoretic norms are irrational or not.

At about the same time, Simon (1956, 1982) elaborated his "bounded rationality" framework. He argued that human judgment and reasoning are limited by their bounds on cognitive capacity, the time available to gather and process information, and the information available. Humans therefore adopt "satisficing" solutions to problems instead of optimal ones whose requirements on cognitive capacity, time, and information may exceed those limits. Unlike exponents of the heuristics and biases school, Simon did not regard satisficing strategies as irrational or erroneous, but instead adaptive.

Gigerenzer and his colleagues expanded Simon's ideas into what some call the "fast and frugal heuristics" school (e.g., Gigerenzer and Selten 2001). Contrary to other theorists' interpretations of Simon's ideas as suggesting that human judgments and decisions are suboptimal, Gigerenzer and associates argued that many of the heuristics people employ do not merely economize on time and cognitive load. They also perform as well as or better than normative strategies by exploiting structure in real environments (e.g., Gigerenzer et al. 1999). Some heuristics that are fallacies in the casino are effective in real although uncertain environments (e.g., the gambler's fallacy, as shown in Smithson 1997).

Finally, other researchers have investigated the possibility that people think and act as if there are kinds of uncertainty distinct from probability. A classic paper by Ellsberg (1961) presented experimental demonstrations that people are influenced in allegedly counter-normative ways by ambiguity, the extent to which a probability is imprecisely specified. He took the position that the norms proposed by Bayesians were inadequate, and that it is reasonable to be influenced by imprecision.

His findings stimulated extensive research and debate over whether "ambiguity aversion" is irrational or not (see Camerer and Weber 1992 for a review). Smithson (1999) reported experiments showing that people also are influenced by uncertainty arising from conflicting information from equally credible sources. Cabantous (2007) replicated and extended Smithson's findings on a sample of professional insurers, and Smithson (2013) developed and tested models of the joint influence of conflict and ambiguity on perceptions of uncertainty. These efforts in psychology and behavioral economics have been paralleled by the creation of alternative formal uncertainty frameworks such as fuzzy logic (Zadeh 1965), belief functions (Shafer 1976), and imprecise probability theories (e.g., Borel 1943/1962, Kyburg 1961, Smith 1961, and Walley 1991). The normative status of these frameworks and whether they are reducible to probability after all are vigorously debated.

23.2 Probability Weighting Functions

There is a large body of empirical and theoretical work on subjective probability judgments that considers them in terms of weighting functions. A widely-accepted account may be summarized in part by saying that people tend to over-weight small and under-weight large probabilities (Kahneman and Tversky 1979), although this account nevertheless is disputed. Note that this claim does not imply that people are under- or over-estimating the probabilities, but instead differentially weighting them when using them for decisions.

Rank-dependent expected utility theory (e.g., Quiggin 1993) reconfigures the notion of a probability weighting function by applying it to a cumulative distribution whose ordering is determined by outcome preferences. Cumulative prospect theory (Tversky and Kahneman 1992) posits separate weighting functions for gains and losses, on the grounds that people are loss-averse in the sense that they are more pessimistic about probabilities of losses and weigh losses more heavily than gains.

Two psychological influences have been offered to explain the properties of probability weighting functions. The hypothesis proposed in Tversky and Kahneman (1992) is "diminishing sensitivity" to changes that occur further away from the reference-points of 0 and 1. This would account for the inverse-S shape, or curvature, of the weighting function typically found in empirical studies (e.g., Camerer and Ho 1994). For instance, a change from .05 to .10 is seen as more significant than a change from .30 to .35, but a change from .65 to .70 is viewed as less significant than a change from .90 to .95.

Gonzales and Wu (1999) added the notion of "attractiveness" of a gamble to account for the elevation of the weighting curve. The magnitude of consequences affects both the location of the inflection-point of the curve and its elevation. Large gains tend to move the inflection point to the left and large losses move it to the right (e.g., Etchart-Vincent 2004).

23.3 Probabilities from Experience versus Description

Until recently the psychology of probability judgments rested chiefly on evidence from judgments made about descriptions of populations of events or event outcomes provided by experimenters. Little attention was paid to judgments made on the basis of experience, with participants sampling events randomly generated from populations with specified event probabilities (not known by the participants) and estimating probabilities on that basis. An experiment reported in Hertwig, Barron, Weber, and Erev (2004) indicated that probability judgments based on experience differ importantly from those based on description. Their primary finding was that in contrast to the view that people overestimate the probabilities of rare events, experience-based judgments *underestimate* the probabilities of such events.

Hertwig et al. called for two theories of probability judgment, one each for judgments based on descriptions and those based on experience. They proposed two explanations account for under-weighting the probability of rare events under experience: small samples and recency effects (i.e., a tendency to give greater weight to the most recent experiences). They noted that unlike humans, other animal species must entirely base their probability judgments on experience. They cited a study of foraging decisions made by bees (Real 1991), which concluded that the bees under-weight rare instances of food and over-weight common instances because, among other factors, bees' samples from foraging experiences are truncated owing to memory constraints.

However, Fox and Hadar (2006) pointed out that Hertwig et al.'s participants showed little or no under-weighting of rare events when their judgments were compared to the actual sample probabilities (i.e., the relative frequencies obtained in the participant's sample) instead of the parent population probabilities. They concluded that the apparent

under-weighting phenomenon is explicable entirely in terms of sampling error and not any psychological influences, a conclusion subsequently supported by others (e.g., Rakow, Demes, and Newell 2008). Researchers responded by attempting to alter the experienced samples so that they more closely match the population probabilities. One effective approach has been to describe the samples seen by each participant in an experience condition to a yoked "partner" in the description condition (Rakow et al., 2008). In a second approach Camilleri and Newell (2011) implemented an algorithm providing small corrections to the sequence of observations to bring the observed payoffs more in line with their objective counterparts. With both methods the description-experience gap largely disappears. Nonetheless, the debate continues over whether this gap is purely due to bias from sampling variability. Rakow and Newell (2010) have recently urged researchers to study the unique contributions that description and experience make, *in combination*, to probability judgments and decisions based on them. Most recently Camilleri and Newell (2011: p. 383) have argued that "repeated and consequential choice... appears to be the crucial element for under-weighting to occur in the absence of sampling bias."

23.4 Probability Judgment Heuristics

Numerous heuristics have been hypothesized to explain people's probability judgments, especially those deviating from the prescriptions of probability theories. We shall review four of these. The question of what could explain biases in probability judgments was first addressed systematically by Kahneman and Tversky (Kahneman and Tversky 1972, 1973; Tversky and Kahneman 1974). They proposed three judgment heuristics to account for biases: anchoring-and-adjustment, representativeness and availability.

The anchoring-and-adjustment heuristic involves two claims. First, people are suggestible to arbitrary numerical estimates when engaged in a numerical estimation task. In an experiment demonstrating anchoring, a wheel yielding a number between 1 and 100 was spun in front of participants, who were then asked to estimate what percentage of UN countries were in Africa. The median estimate for a group whose starting number was 10 was that 25 percent of UN countries were African; the median estimate for the group whose starting number was 65 was nearly twice that: 45 percent. The wheel number biased their estimates even though they knew that the number was irrelevant (Tversky and Kahneman 1974).

The second claim is that people adjust away from an initial estimate insufficiently when presented with evidence. A version of this claim specific to probability judgments is older than the anchoring and adjustment hypothesis (Phillips and Edwards 1966) and is sometimes called "conservatism," i.e., that the difference between a person's prior and posterior probability estimates is less than that prescribed by Bayes' theorem. The mechanisms that underpin anchoring and adjustment have been debated (e.g., Epley and Gilovich 2004). Initially, the bias towards initial anchors was ascribed to insufficient adjustment, as in conservatism, but later research suggested alternative influences, such as confirmation bias.

The representativeness heuristic is that people judge probability via similarity or prototypicality. That is, they ask themselves how similar an item is to a typical member of

the category in order to assess the probability that the item belongs to that category. This heuristic has been used to explain people's neglect of relevant base-rates when judging single-event probabilities, and tendency to produce estimates too far from the mean when making predictions based on imperfect predictors. Kahneman and Tversky (1973) demonstrated this heuristic by showing that one group of subjects' judgments of how likely a student was to be majoring in each of nine fields closely matched the judgments made by another group of how similar that student was to the typical student in each of those fields.

The availability heuristic is that people judge the probability of an event by how readily previous occurrences come to mind. It is easier to think of English words beginning with K than words that have K as the third letter, so most people judge that words of the first kind are the more numerous, when in fact the latter is more frequent. The availability heuristic has been used to explain the Catch-All Underestimation Bias (CAUB, in Fischhoff, Slovic, and Lichtenstein 1978), whereby if event categories are combined under a single super-set then the probability people assign to the super-set typically is less than the sum of the probabilities people assign to its component categories (see also Tversky and Koehler 1994).

A more recent example of a probability judgment heuristic is the affect heuristic, a tendency to assess probabilities of outcomes based on how one feels about those outcomes. According to Slovic and Peters (2006), people judge an outcome as less risky if they are favorably disposed towards it, and they consider it more risky if their feelings about it are negative. For example, people fear radiation from nuclear power plants more than they fear radiation from medical X-rays, whereas for most people X-rays pose the greater threat.

These heuristics have stimulated considerable research but also serious criticism. For example, Gigerenzer (1996: p. 592) averred that the representativeness, availability, and anchoring heuristics are unfalsifiable "…because, post hoc, one of them can be fitted to almost any experimental result. For example, base-rate neglect is commonly attributed to representativeness. However, the opposite result, over-weighting of base-rates (conservatism), is as easily 'explained' by saying the process is anchoring…" Even a more sympathetic author such as Bar-Hillel (1984) admitted that the counter-normative neglect of three factors (base-rate, sample size, and prediction error) in probability judgments is distinct from the representativeness heuristic and not necessarily explained by it. Attempts to develop more extensively elaborated models of the cognitive processes underlying probability judgments have met with mixed success (see below).

23.5 Partition Dependence and Additivity

The aforementioned CAUB phenomenon has been explained not only by the availability heuristic, but also by partition dependence. On grounds of insufficient reason, a probability of $1/K$ is assigned to K mutually exclusive possible events when nothing is known about the likelihood of those events. Fox and Rottenstreich (2003) presented evidence that subjective probability judgments are typically biased towards this ignorance prior. Thus, probability judgments influenced by the ignorance prior are partition-dependent. Fox and Clemen

(2005) found evidence that this dependence decreases as domain knowledge increases, but that even experts in decision analysis are susceptible to it.

Partition dependence poses problems for probability assignments in two respects. First, it may be unjustified because there is a normatively correct partition. For instance, Fox and Rottenstreich (2003) posed the question of how likely Sunday is to be the hottest day of the week. The principle of insufficient reason would suggest that 1/7 is the correct answer, so their demonstration that people can be induced to partition the events into just two possibilities (Sunday is or is not the hottest day) and therefore assign a probability of 1/2 indicates that those people are anchoring onto an incorrect partition.

The second difficulty arises when there is no normatively correct partition or the sample space is ambiguous. Consider a bag containing 10,000 marbles whose colors are completely unknown to us. How should we use the principle of insufficient reason to judge the probability of drawing a red marble from this bag? Also, consider a scenario in which we are told that Chris, James, and Pat are applying to an engineering firm and are then asked to estimate the probability that each of them is hired by the firm. It is unclear whether there is only one position available or multiple positions, or whether these three are the only applicants. Thus, equally defensible partitions could yield ignorance priors of 1/3 each, 1/2 each, or p each, where p is any rational number such that $0 < p \leq 1$. The assignment $p = r/k$, for $k \geq r$, is consistent with assuming there are r positions and k applicants.

Smithson et al. (2010) used this scenario in experiments to show that a simple cue asking people to nominate which candidate has the highest probability of being hired induces more people to constrain their probabilities to strict additivity (i.e., summing to 1 and thus anchoring on a 3-fold partition). Moreover, they found that Japanese respondents were less likely to insist on additivity than Australians. Gelfand and Christakopoulou (1999) report that Americans are more likely to assume a fixed pie (zero-sum) in negotiations than Greeks, which they attribute to the individualism in the former and collectivism in the latter culture. Smithson et al. (2010) found indirect evidence that this could account for the Japanese–Australian difference.

A striking example of strict additivity where none is required comes from a study by Sopena (2009). Australian participants were presented with descriptions of migrants from Syria and Canada and were asked to make judgments regarding the degree to which they considered these targets as prototypically Australian and non-Australian, on a rating scale from 0 to 20 for each judgment. Targets were described as either highly threatening or non-threatening on a variety of issues (e.g., fundamentalist religious orientation). Degrees of membership in the sets of Australians and Syrians or Canadians need not sum to 20 (the sets can overlap). However, Smithson's reanalysis of Sopena's data revealed that respondents' ratings summed exactly to 20 more often for high-threat and Syrian targets. These findings suggest that additivity of group membership increases under perceived threat or when the target is an outgroup member, despite the absence of any normative justification for additivity.

Another influence is a predisposing cognitive bias or stereotype that bolsters a belief that a resource is zero-sum. In Meegan's (2010) experiments participants were undergraduates at a university that does not "grade on the curve," but instead awards grades on the basis of fixed criteria rather than fixed quotas. Thus, grades at that university are not zero-sum. Nevertheless, when participants were asked to predict the grade of a student after they had been shown a skewed distribution of grades already assigned in the same class where the

majority were high grades, they predicted a lower grade than participants who had been shown a symmetric distribution. However, the effect disappeared when participants viewed a grade distribution where most grades were low. Meegan concluded that for desirable outcomes people tend towards zero-sum thinking when presented with others' gains but not when presented with their losses.

23.6 Calibration and Overconfidence

Even if human probability judgments are not accurate, they may still be well-calibrated in the sense that they are neither consistently too high nor too low. In a classic study, Murphy and Winkler (1977) examined the calibration of 25,000 weather forecasts made over a four-year period which included probability estimates of rain, snow, and other weather events and established that their calibration was very good. However, a review by Lichtenstein, Fischhoff, and Phillips (1982) of the empirical literature on lay confidence judgments indicated that people tend to be over-confident when their confidence is generally high and under-confident when their confidence is generally low.

Likewise, a large empirical literature on subjective confidence–interval estimation tasks suggests that people are badly-calibrated (overconfident) in the sense that they construct intervals that are too narrow for the confidence level nominated (e.g., Alpert and Raiffa 1982; Klayman, Soll, Gonzalez-Vallejo, and Barlas 1999). Nor is this confined to lay people. In a study of experts' judgmental estimates (Russo and Schoemaker 1992) in which business managers estimated 90% confidence-intervals for uncertain quantities in their areas of expertise (e.g., petroleum, banking, etc), the hit rates obtained in various samples of managers ranged from 38% to 58%. These are performance levels similar to those typically found in studies of lay people, which indicates that domain expertise does not necessarily confer calibration when it comes to subjective probability estimation.

Yaniv and Foster (1995, 1997) suggested that judgments and evaluations of subjective interval estimates are the product of two competing objectives: accuracy and informativeness. They hypothesized that the patterns of preference ranking for judgments support a simple trade-off model between precision (width) of interval estimates and absolute error which they characterized by the error-to-precision ratio. Both papers presented arguments and evidence that people tend to prefer narrow but inaccurate interval estimates over wide but accurate ones, i.e., they value informativeness more highly than accuracy. For instance, study 3 in Yaniv and Foster (1995) asked participants to choose between two estimates, (A) [140,150] and (B) [50,300]. They were told that the correct answer was 159. A large majority (90%) of the respondents preferred estimate A over B, although only the latter interval includes the correct answer.

Some researchers found that the format for eliciting interval estimates influences overconfidence. Soll and Klayman (2004) compared overconfidence in interval estimates using three elicitation methods: Range, two-point intervals and three-point intervals. The range method simply asks for, say, a 90% subjective confidence interval. The two-point method asks for a lower limit with a 95% chance of falling below the true value and then an upper limit with a 95% chance of falling above it. The three-point method adds a "midpoint"

estimate with a 50% criterion. They found the least overconfidence where the three-point method was used and the greatest with the range method. Their explanation was that the two- and three-point methods encourage people to sample their knowledge at least twice, whereas the range method is treated by most people as a single sample from their knowledge base. Other aspects of the task that have been investigated include the extremity of the confidence criterion and the nature of the scale used for elicitation (e.g., the study of graininess effects in Yaniv and Foster 1997). For example, Garthwaite and O'Hagan (2000) found that tertiles—the 1/3 and 2/3 quantiles—yielded better calibration than more extreme confidence levels.

A major puzzle in this area was the repeated finding that while people are overconfident when they construct intervals, they are reasonably well-calibrated when asked to assign probabilities to two-alternative questions with the same estimation targets (Klayman et al. 1999). An example of such a task is asking the respondent whether the population of Thailand exceeds 25 million (yes or no) and then asking for the subjective probability that her answer is correct.

A breakthrough came when Winman, Hansson, and Juslin (2004) revised the two-alternative question format for probability judgments about interval estimates provided to the respondent (e.g., estimating the probability that Thailand's population is between 25 million and 35 million). Comparing these judgments with the intervals elicited from respondents with a fixed confidence criterion (e.g., a 90% subjective confidence interval for Thailand's population), they found that overconfidence was nearly absent in the intervals provided but, as always, high in the elicited intervals. These findings have since been replicated in most, but not all, comparison experiments of this kind (O'Hagan et al. 2006).

Juslin, Winman, and Hansson (2007) partly accounted for these and related findings by noting that while a sample proportion is an unbiased estimate of the true probability, the sample confidence interval coverage-rate is upwardly biased. They hypothesized that people, as "naïve intuitive statisticians," are relatively accurate in sampling their own knowledge but treat all sample estimates as if they are unbiased. The implication is that subjective probability judgments of intervals provided to respondents are better calibrated than intervals produced by respondents to match a coverage-rate.

23.7 Conjunction and Conditional Probability Fallacies

Judgments of compound-event and conditional probabilities have been studied extensively. For compound events, most attention has been focused on the so-called *conjunction fallacy*, which is the tendency to violate the conjunction rule that $P(A\&B) \leq \min(P(A), P(B))$. Evidence for these violations first was reported by Kahneman, Slovic, and Tversky (1982) and Tversky and Kahneman (1983). Numerous replications followed, and although some scholars questioned whether these really constituted a fallacy (e.g., Wolford, Taylor, and Beck 1990), the general consensus has been that the conjunction fallacy is a consistent bias in human reasoning about probabilities.

Nevertheless, a somewhat ironic comparison can be made with Osherson and Smith's (1981) critique of fuzzy set theory's adherence to the rule that the degree of membership in the conjunction of two categories cannot exceed membership in either component. They presented counter-examples, such as a guppy, which is more prototypical of "pet fish" than either of "pet" or of "fish". Apparently, what is a fallacy for intuitive probability judgments is not for judgments of membership; and whereas mathematics overrules human intuition in probability, human intuition overrules mathematics in categorization.

Several competing explanations for the conjunction fallacy have been advanced. The earliest explanation linked it to the representativeness heuristic (Tversky and Kahneman 1983). In their famous example, they argued that when people judge the probability that Linda is a "bank teller" or "feminist bank teller", they judge her description to be more similar to "feminist bank teller" than to "bank teller" and therefore assign a higher probability to the conjunction than to the "bank teller" component.

Another explanation was the signed summation model (Yates and Carlson 1986), in which low-probability events are assigned a negative value on a subjective scale and high-probability events a positive value, the idea being that a conjunctive event is judged by the sum of these signed values so that the sum of negative- and positive-scored events will be greater than the negative score of the first event. A third explanation (Thüring and Jungermann 1990) was that the subjective probability of a conjunction typically over-weights the smaller component so that the fallacy arises only when the two component probabilities differ substantially. A fourth account arose from support theory (Tversky and Koehler 1994), which predicts that because probabilities are judged by the number of supporting events, a conjunction will be judged to have greater probability than either of its components.

Agnoli and Krantz (1989) were the first to suggest that competing heuristics might be involved in judgments of conjunctive probabilities and that their dominance would be context-sensitive. Fisk and Pidgeon (1996, 1997) added to this proposal the notion of process-based reasoning, whereby people estimate these probabilities in two stages by anchoring on the smaller component and then adjusting away from that. However, the greatest restraint on the rate at which people commit this fallacy was achieved by Fiedler (1988) who refashioned the problem into judgments about relative frequency. Fiedler asked participants to estimate how many of 100 people "who are like Linda" would fall into the conjunctive and constituent categories. His results were a reduction from 91% to 22% of participants committing the conjunction fallacy.

A related and widely-known fallacy is when people are asked for P(H|E), the probability that an hypothesis H is true given evidence E, when given information about P(E) and the base-rate P(H), they often respond with P(E|H) instead. Numerous examples of this confusion of inverse conditional probabilities have been found in medicine and law, but the most famous example is Tversky and Kahneman's (1982) taxicab problem. Participants were told that 85% of the cabs in a city are Green and 15% Blue; a witness to a hit-and-run accident at night claims the offending cab was Blue; and the witness has been found to identify each of the two colors correctly at nighttime 80% of the time. They were then asked for the probability that the cab in the accident was Blue. Many responded with 80%, i.e., P(E|H).

One implication Tversky and Kahneman drew from responses to the taxicab problem was that participants ignored the base-rate information (i.e., the 15% Blue and 85%

Green cabs). A more direct demonstration of *base-rate neglect*, the tendency to ignore or down-weight base-rate information, was Kahneman and Tversky's (1973) study in which participants were told that a brief description of a professional had been randomly drawn from 100 descriptions, 30 of engineers and 70 of lawyers. When asked the probability that the description was of, say, an engineer their ratings were influenced by how similar they thought the description was to their stereotype of the specified profession. Even an "information-free" description yielded a probability of .5 from most subjects.

As with the conjunction fallacy, several explanations have been proposed for base-rate neglect. In addition to appeals to representativeness and stereotyping, it has been hypothesized that people consider base-rate information irrelevant (Cohen 1981) or that they confuse P(E|H) with P(H|E) (Eddy 1982). Hamm (1993) found that some subjects' responses are consistent with the latter explanation and others with the proposal that people interpolate between the base-rate and 1.0.

Also as with the conjunction fallacy, revising the format of the problem into a frequency format greatly increased the percentage of correct answers. Cosmides and Tooby (1996) experimentally demonstrated this with a problem from a study by Casscells, Schoenberger, and Graboys (1978): If a diagnostic test for a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the probability that a person with a positive test result actually has the disease (if no other information is available)? In the Casscells et al. sample of Harvard Medical School students and staff, only 18% gave the correct answer (.02), and Cosmides and Tooby reported a correct response rate of 12% under the same conditions. Rephrasing the problem in a frequentist way increased that rate to 76% and adding a visual response format (a 10x10 grid with each square representing one person randomly sampled from the population) increased it to 92%.

23.8 Communication of Probabilities

In addition to human probability judgments, the communication of probabilistic risk information has been studied extensively (see Budescu and Wallsten 1995 for a thorough review). This research began with the notion that if verbal probability expressions, such as "likely" or "improbable", have an agreed-upon numerical translation then both elicitation and communication tasks can be simplified by referring to a "dictionary" of these expressions. Indeed, several studies (e.g., Brun and Teigen 1988, Erev and Cohen 1990, Wallsten et al. 1993) have reported a widespread preference among people for communicating uncertainties by using verbal expressions rather than numbers. This literature also has debated the question of whether probabilities are better communicated via numbers than by words. There is a long history of attempts to translate verbal probability expressions into numerical form and debates over whether the results are sufficiently reliable and consensual.

In the earliest studies people were asked to nominate single numbers to represent probability expressions (PEs). These studies reported reasonably high intra-subjective reliability (e.g., Lichtenstein and Newman 1967; Beyth-Marom 1982; Budescu and Wallsten 1985; Budescu, Weinberg, and Wallsten 1988) and reliable aggregate means (Simpson 1944; Reagan, Mosteller, and Youtz 1989). However, the same research also revealed

considerable inter-subjective variability and overlap among phrases (Stone and Johnson 1959; Lichtenstein and Newman 1967; Beyth-Marom 1982; and Boettcher 1995). Budescu and Wallsten (1985) argued that PEs may lead to ordinal confusion in communication, and Budescu, Weinberg and Wallsten (1988) provided evidence that people vary widely in the PEs they regularly use.

One reasonable interpretation of these findings is that they are symptomatic of vagueness or fuzziness and not just individual differences. Wallsten et al. (1986) established an experimental paradigm in which subjects constructed fuzzy membership functions over the unit interval to translate PEs into numerical terms (see the material on interval evaluation below). Their approach was among the earliest to systematically explore the connection between PEs and *imprecise probabilities*, which include probability intervals and sets of probabilities (see Cozman, (14) in this volume). Kent (1964) anticipated this idea by proposing probability intervals as translations of a set of PEs he hoped would be adopted by American intelligence operatives. However, although the British intelligence community eventually adopted this approach, the American intelligence community did not (Kesselman 2008). Translations of PEs into numerical imprecise probabilities seem likely to succeed only in small communities of experts who can agree on nomenclature and, of course, the translation itself.

A recent attempt to impose such a translation on the public at large is the Intergovernmental Panel on Climate Change fourth report (IPCC 2007), which presented a collection of nested intervals corresponding with a set of PEs used throughout their report. Budescu et al. (2009, 2014) found that people's estimates of the probabilities corresponding to the PEs in IPCC report sentences were more regressive (towards the middle of the unit interval) than intended by the IPCC authors, an effect that was only partially reduced by embedding an explicit numerical interval in each sentence.

Interpretations of PEs also have proven to be context-dependent. Weber and Hilton (1990) investigated outcome severity in the context of medical scenarios and found that PEs were mapped to higher probabilities when associated with more severe events. Likewise, PEs appear to be vulnerable to partition priming effects (Tsao and Wallsten 1994). A related literature discusses partition effects for imprecise probabilities. Walley (1991) argued that while partition dependence poses an unavoidable problem for likelihood judgments that yield a single probability, imprecise probability judgments need not depend on the state-space partition. For instance, under complete ignorance Walley recommended that we assign a lower probability of 0 and an upper probability of 1 to every event, no matter what the partition is. Whether or when imprecise probability judgments and PEs are or should be partition-dependent is an open question. Smithson and Segale (2009) compared elicited probability intervals with elicited precise probabilities and found that the locations of the intervals were as much influenced by partition priming as were those of precise probabilities. However, they also found that some respondents constructed wider intervals when primed with an incorrect partition, indicating that respondents still bore the correct partition in mind.

Some researchers have highlighted relevant differences between meanings or usages in natural versus formal language. For instance, negation was found to be asymmetric in its effects, so that "unlikely" is not subjectively equivalent to 1, "likely". More specifically, PEs have been found to be more inherently "directional" than numbers (Teigen and Brun 1995) and positive PEs tend to be applied to a wider range of numerical probabilities and

outcomes than negative PEs. Smithson et al. (2012) reported more regressive responses and less inter-subjective consensus for negative than positive PEs when respondents were asked to translate them from sentences used in the IPCC fourth report (2007). Outcome valence also may affect the interpretation of PEs, although this issue has not been extensively investigated. Mullet and Rivet (1991) compared probabilities assigned to 24 French expressions used in predictions of children's chances of passing or failing a test. On average, a positive context induced higher probabilities for a given phrase.

All told, a clear implication of this line of research is that PEs are not an effective way to communicate probabilities and should be avoided. However, several studies have found that people are no less Bayesian (Rapoport et al. 1990), no more over-confident (Wallsten et al. 1993), and no worse at betting, bidding and decision-making (Budescu and Wallsten 1990, Gonzalez-Vallejo et al. 1994) when they use PEs than when they use numbers. Wallsten et al. (1988) proposed a potential resolution of this apparent contradiction by hypothesizing that for decisional purposes people "resolve" the vagueness in a PE by focusing on a numerical probability that they consider prototypical of the PE. Thus, the case against PEs is not entirely convincing.

Finally, relatively little research has been conducted on motivational influences on the communication of uncertainty, despite the fact that normative concerns about bluffing and bid-ask spreads (the difference between the buyer's highest price and the seller's lowest price) extend back to at least Borel (1924/1964: p. 58) and produced the literature on scoring rules for eliciting "honest" probability judgments (e.g., Brier 1950). Schweitzer and Hsee (2002) experimentally demonstrated that motivational factors exert greater influence on estimations elicited from participants under high than under low uncertainty conditions. They argued that greater uncertainty creates leeway for decision-makers to justify extreme (and self-serving) claims for themselves. From a normative viewpoint, Seidenfeld, Schervish, and Kadane (2011) provided a formal argument that there is no single real-valued scoring rule that can play the same regulatory role for imprecise probabilities as Brier scores can for precise probabilities.

23.9 Normative Fundamentalism?

Because so much research on human probability judgments has compared them with some version of a Bayesian framework which is taken as the benchmark of rationality, we shall conclude by briefly noting three lines of criticism that have been leveled at this approach. The more moderate criticism has been that comparisons with Bayesian prescriptions do not sufficiently address questions of how people construct their judgments. Elqayam and Evans (2011) claim that an "ought-is" distinction biases research by restricting attention to normative correlates and neglecting philosophically significant questions that lack a clear standard for normative judgment. Earlier, Gigerenzer (1993) called for greater attention to the mental models and cognitive processes involved in probability judgments and less of a focus on errors and biases. Indeed, recent trends in the literature on probability judgment have been towards greater emphasis on modeling cognitive processes.

A more radical line of criticism has been that the normative standards are inappropriate or miss-specified. For instance, Gigerenzer (1991) and Teigen (1994), among others,

have argued that distinctions such as the one between single-event and relative-frequency probabilities yield important disputes about normative standards for probability judgments. Gigerenzer points to prominent proponents of the frequentist school of probability theory, who regard the concept of a probability of a unique event as meaningless. He also advocates comparing "apples with apples," for instance, comparing objective relative frequencies with people's estimates of relative frequencies instead of with their confidence judgments. Likewise, Teigen claims that at least some of the "deviations" from Bayesian prescriptions can be explained by defensible reasoning depending on whether probabilities are being judged on the basis of base-rates (relative frequencies), internal mental states, dispositions, or degrees of plausibility. Crupi, Fitelson, and Temtori (2008) argue that experimentally observed fallacious probability judgments in conjunction problems may be guided by sound assessments of confirmation relations (as in Bayesian confirmation theory).

The third critique has been alluded to earlier, namely that there may be other kinds of uncertainty such as ambiguity or conflict, and when people are influenced by these they are not behaving irrationally. Arguments defending ambiguity aversion include aversion to missing but obtainable information (Ritov and Baron 1990) and sensitivity to variability of outcomes (e.g., Rode et al. 1999). Outcome variance relative to outcome magnitude is an effective predictor of responses to risk in human and non-human species (Weber, Shafir, and Blais 2004). This finding has been defended via optimal foraging theory (Caraco 1981), whereby animals choose the option with lower variance if the mean caloric payoff exceeds current need but choose the higher-variance option if the mean payoff falls below current need. Similar arguments may be extended to conflict aversion, along with the question of trustworthiness of the conflicting message sources when they are supposedly based on the same information (Smithson 1999).

These critiques and the research and theory development stemming from them have deepened our understanding of the psychological aspects of probability judgment, and judgments of risk in general. They also have stimulated and contributed to the ongoing debates about the normative status of precise probabilities and candidates for other kinds of uncertainty.

References

Agnoli, F. and Krantz, D. H. (1989) Suppressing Natural Heuristics by Formal Instruction: The Case of the Conjunction Fallacy. *Cognitive Psychology.* 21. pp. 515–50.

Alpert, M. and Raiffa, H. (1982) A Progress Report on the Training of Probability Assessors. In Kahneman, D., Slovic, P., and Tversky, A. (eds.) *Judgment under Uncertainty: Heuristics and Biases*. pp. 294–305. New York, NY: Cambridge University Press.

Bar-Hillel, M. (1984) Representativeness and Fallacies of Probability Judgment. *Acta Psychologica*. 55. pp. 91–107.

Beyth-Marom, R. (1982) How Probable is Probable? A Numerical Translation of Verbal Probability Expressions. *Journal of Forecasting*. 1. pp. 257–69.

Boettcher, W.A. (1995) Context, Methods, Numbers, and Words: Prospect Theory in International Relations. *Journal of Conflict Resolution*. 39. pp. 561–83.

Borel, E. (1921/1953) The Theory of Play and Integral Equations with Skew-symmetric Kernels. Translated from the French by L. J. Savage. *Econometrica*. 21. pp. 97–100.

(Originally published in *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences*. 173. pp. 1304–8.)

Borel, E. (1924/1964) Apropos of a Treatise on Probability. Translated from the French by H. E. Smokler. In Kyburg, H. E., Jr. and Smokler, H. E. (eds.) *Studies in Subjective Probability*. pp. 45–60. New York: Wiley.

Borel, E. (1943/1962) *Probabilities and Life*. Translated from the French by M. Baudin. New York, NY: Dover.

Brier, G. W. (1950) Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*. 78. pp. 1–3.

Brun, W. and Teigen, K. H. (1988) Verbal Probabilities: Ambiguous, Context-Dependent, or Both? *Organizational Behavior and Human Decision Processes*. 41. pp. 390–404.

Budescu, D. V., Broomell, S. B., and Por, H. H. (2009) Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change. *Psychological Science*. 20. pp. 299–308.

Budescu, D. V., Por, H.-H., Broomell, S. B., and Smithson, M. (2014) The Interpretation of IPCC Probabilistic Statements around the World. *Nature Climate Change*. 4. pp. 508–12.

Budescu, D. V. and Wallsten, T. S. (1985) Consistency in Interpretation of Probabilistic Phrases. *Organizational Behavior and Human Decision Processes*. 36. pp. 391–405.

Budescu, D. V. and Wallsten, T. S. (1990) Dyadic Decisions with Verbal and Numerical Probabilities. *Organizational Behavior and Human Decision Processes*. 46. pp. 240–63.

Budescu, D. V. and Wallsten, T. S. (1995) Processing Linguistic Probabilities: General Principles and Empirical Evidence. In Busemeyer, J., Hastie, R., and Medin, D. L. (eds.) *Decision Making from a Cognitive Perspective*. pp. 275–318. San Diego: Academic Press.

Budescu, D. V., Weinberg, S., and Wallsten, T. S. (1988) Decisions Based on Numerically and Verbally Expressed Uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*. 14. pp. 281–94.

Cabantous, L. (2007) Ambiguity Aversion in the Field of Insurance: Insurers' Attitude to Imprecise and Conflicting Probability Estimates. *Theory and Decision*. 62. pp. 219–40.

Camerer, C. F. and Ho, T. H. (1994) Violations of the Betweenness Axiom and Nonlinearity in Probability. *Journal of Risk and Uncertainty*. 8. pp. 167–96.

Camerer, C. and Weber, M. (1992) Recent Developments in Modeling Preferences: Uncertainty and Ambiguity. *Journal of Risk and Uncertainty*. 5. pp. 325–70.

Camilleri, A. R. and Newell, B. R. (2011) When and Why Rare Events are Underweighted: A Direct Comparison of the Sampling, Partial Feedback, Full Feedback and Description Choice Paradigms. *Psychonomic Bulletin & Review*. 18. pp. 377–84.

Caraco, T. (1981) Energy Budget, Risk, and Foraging Preferences in Dark-Eyed Juncos. *Behavioral Ecological Sociobiology*. 8. pp. 820–30.

Casscells, W., Schoenberger, A., and Graboys, T. B. (1978) Interpretation by Physicians of Clinical Laboratory Results. *New England Journal of Medicine*. 299. pp. 999–1001.

Cohen, J. (1954) Conjecture and Risk. *The Advancement of Science Reports of the British Association*. 11. pp. 333–9.

Cohen, J. (1960) *Chance, Skill and Luck*. Harmondsworth: Penguin.

Cohen, J. (1964) *Behaviour in Uncertainty*. London: Allen & Unwin.

Cohen, L. J. (1981) Can Human Irrationality be Experimentally Demonstrated? *Behavioral and Brain Sciences*. 4. pp. 317–70.

Cosmides, L. and Tooby, J. (1996) Are Humans Good Intuitive Statisticians After All? Rethinking some Conclusions from the Literature on Judgment under Uncertainty. *Cognition*. 58. pp. 1–73.

Cozman, F. (2016) Imprecise Probabilities. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. pp. 297–314. Oxford: Oxford University Press.

Crupi, V., Fitelson, B., and Temtori, K. (2008) Probability, Confirmation and the Conjunction Fallacy. *Thinking and Reasoning*. 14. pp. 182–99.

de Finetti, B. (1937/1964) Foresight: Its Logical Laws, its Subjective Sources. Translated by H.E. Kyburg. In Kyburg, H. E, Jr., and Smokler, H. E, (eds.) *Studies in Subjective Probability*. pp. 93–158. New York: Wiley.

Eddy, D. M. (1982) Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities. In Kahneman, D., Slovic, P., and Tversky, A. (eds.) *Judgment under Uncertainty: Heuristics and Biases*. pp. 249–67. New York: Cambridge University Press.

Edwards, W. (1954) The Theory of Decision Making. *Psychological Review*. 51. pp. 380–417.

Edwards, W. (1961) Behavioral Decision Theory. *Annual Review of Psychology*. 12. pp. 473–98.

Ellsberg, D. (1961) Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*. 75. pp. 643–69.

Elqayam, S. and Evans, J. S. (2011) Subtracting "Ought" from "Is": Descriptivism versus Normativism in the Study of Human Thinking. *Behavioral and Brain Sciences*. 34. pp. 233–90.

Epley, N. and Gilovich, T. (2004) Are Adjustments Insufficient? *Personality and Social Psychology Bulletin*. 30. pp. 447–60.

Erev, I. and Cohen, B. L. (1990) Verbal versus Numerical Probabilities: Efficiency, Biases, and the Preference Paradox. *Organizational Behavior and Human Decision Processes*. 45. pp. 1–18.

Etchart-Vincent, N. (2004) Is Probability Weighting Sensitive to the Magnitude of Consequences? An Experimental Investigation on Losses. *Journal of Risk and Uncertainty*. 28. pp. 217–35.

Fiedler, K. (1988) The Dependence of the Conjunction Fallacy on Subtle Linguistic Factors. *Psychological Research*. 50. pp. 123–9.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978) Fault Trees: Sensitivity of Estimated Failure Probabilities to Problem Representation. *Journal of Experimental Psychology: Human Perception Performance*. 4. pp. 330–44.

Fisk, J. E. and Pidgeon, N. F. (1996) Component Probabilities and the Conjunction Fallacy: Resolving Signed Summation and the Low Probability Model in a Contingent Approach. *Acta Psychologica*. 94. pp. 1–20.

Fisk, J. E. and Pidgeon, N. F. (1997) The Conjunction Fallacy: The Case for the Existence of Competing Heuristic Strategies. *British Journal of Psychology*. 88. pp. 1–27.

Fox, C. R. and Clemen, R. T. (2005) Subjective Probability Assessment in Decision Analysis: Partition Dependence and Bias toward the Ignorance Prior. *Management Science*. 51. pp. 1417–32.

Fox, C. R. and Hadar, L. (2006) "Decisions from experience" = sampling error + prospect theory: Reconsidering Hertwig, Barron, Weber, & Erev (2004). *Judgment and Decision Making*. 1. pp. 159–61.

Fox, C. R. and Rottenstreich, Y. (2003) Partition Priming in Judgment under Uncertainty. *Psychological Science*. 13. pp. 195–200.

Garthwaite, P. H. and O'Hagan, A. (2000) Quantifying Expert Opinion in the UK Water Industry: An Experimental Study. *The Statistician*. 49. pp. 455–77.

Gelfand, M. J. and Christakopoulou, S. (1999) Culture and Negotiator Cognition: Judgment Accuracy and Negotiation Processes in Individualistic and Collectivistic Cultures. *Organizational Behavior and Human Decision Processes*. 79. pp. 248–69.

Gigerenzer, G. (1991) From Tools to Theories: A Heuristic of Discovery in Cognitive Psychology. *Psychological Review*. 98. pp. 254–67.

Gigerenzer, G. (1993) The Bounded Rationality of Probabilistic Mental Models. In Manktelow, K. I. and Over, D. E. (eds.) *Rationality: Psychological and Philosophical Perspectives*. pp. 285–313. London: Routledge.

Gigerenzer, G. (1996) On Narrow Norms and Vague Heuristics: A Reply to Kahneman and Tversky (1996). *Psychological Review*. 103. pp. 592–6.

Gigerenzer, G. and Selten, R. (eds.) (2001) *Bounded Rationality: The Adaptive Toolbox*. Cambridge, MA: MIT Press.

Gigerenzer, G., Todd, P. M., and the ABC Research Group (1999) *Simple Heuristics that Make Us Smart*. London: Oxford University Press.

Gonzales, R. and Wu, G. (1999) On the Shape of the Probability Weighting Function. *Cognitive Psychology*. 8. pp. 129–66.

Gonzalez-Vallejo, C. C., Erev, I., and Wallsten, T. S. (1994) Do Decision Quality and Preference Order Depend on Whether Probabilities are Verbal or Numerical? *American Journal of Psychology*. 107. pp. 157–72.

Hamm, R. M. (1993) Explanations for Common Responses to the Blue/green Cab Probabilistic Inference Word Problem. *Psychological Reports*. 72. 219–42.

Hertwig, R., Barron, G., Weber, E. U., and Erev, I. (2004) Decisions from Experience and the Effect of Rare Events in Risky Choice. *Psychological Science*. 15. pp. 534–9.

Hogarth, R. M. (1980) *Judgment and Choice: The Psychology of Decision*. Chichester: Wiley.

IPPC [Intergovernmental Panel on Climate Change] (2007) *Summary for Policymakers*: Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Available from: <http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-spm.pdf(2007)> [Accessed: May 2010].

Juslin, P., Winman, A., and Hansson, P. (2007) The Naïve Intuitive Statistician: A Naïve Sampling Model of Intuitive Confidence Intervals. *Psychological Review*. 114. pp. 678–703.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982) *Judgment under Uncertainty: Heuristics and Biases*. New York, NY: Cambridge University Press.

Kahneman, D. and Tversky, A. (1972) Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology*. 3. pp. 430–54.

Kahneman, D. and Tversky, A. (1973) On the Psychology of Prediction. *Psychological Review*. 80. pp. 237–51.

Kahneman, D. and Tversky, A. (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*. 47. pp. 263–91.

Kent, S. (1964) Words of Estimative Probability. *Studies in Intelligence*. 8. pp. 49–65.

Kesselman, R. F. (2008) Verbal Probability Expressions in National Intelligence Estimates: A Comprehensive Analysis of Trends from the Fifties through Post 9/11. Unpublished Masters Thesis, Mercyhurst College, PA.

Keynes, J. M. (1921) *A Treatise on Probability*. London: Macmillan.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., and Barlas, S. (1999) Overconfidence: It Depends on How, What, and Whom You Ask. *Organizational Behavior and Human Decision Processes*. 79. pp. 216–47.

Kyburg, H. E. Jr. (1961) *Probability and the Logic of Rational Belief*. Middletown: Wesleyan University Press.

Lichtenstein, S., Fischhoff, B, and Phillips, B. (1982) Calibration of Probabilities: The State of the Art to 1980. In Kahneman, D., Slovic, P., and Tversky, A. (eds.) *Judgment under Uncertainty: Heuristics and Biases*. pp. 306–34. New York, NY: Cambridge University Press.

Lichtenstein, S. and Newman, J. R. (1967) Empirical Scaling of Common Verbal Phrases Associated with Numerical Probabilities. *Psychonomic Science*. 9. pp. 563–4.

Meegan, D. V. (2010) Zero-Sum Bias: Perceived Competition Despite Unlimited Resources. *Frontiers in Psychology: Cognition*. 1. pp. 1–7.

Mullet, E. and Rivet, I. (1991) Comprehension of Verbal Probability Expressions in Children and Adolescents. *Language and Communication*. 11. pp. 217–22

Murphy, A. H. and Winkler, R. L. (1977) Can Weather Forecasters Formulate Reliable Probability Forecasts of Precipitation and Temperature? *National Weather Digest*. 2. pp. 2–9.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. Chichester: Wiley.

Osherson, D. N. and Smith, E. E. (1981) On the Adequacy of Prototype Theory as a Theory of Concepts. *Cognition*. 9. pp. 35–58.

Phillips, L. D. and Edwards, W. (1966) Conservatism in a Simple Probability Inference Task. *Journal of Experimental Psychology*. 72. pp. 346–54.

Quiggin, J. (1993) *Generalized Expected Utility Theory: The Rank Dependent Model*. Boston: Kluwer.

Rakow, T., Demes, K. A., and Newell, B. R. (2008) Biased Samples not Mode of Presentation: Reexamining the Apparent Underweighting of Rare Events in Experience-based Choice. *Organizational Behavior and Human Decision Processes*. 106. pp. 168–79.

Rakow, T. and Newell, B. R. (2010) Degrees of Uncertainty: An Overview and Framework for Future Research on Experience-Based Choice. *Journal of Behavioral Decision Making*. 23. pp. 1–14.

Ramsey, F. P. (1926/1931) Truth and Probability. In Ramsey, F. P. (ed. Braithwaite, R. B.) *The Foundations of Mathematics and Other Logical Essays*. pp. 156–98. London: Routledge and Kegan Paul. (Published posthumously in this form. First published as a paper.)

Rapoport, A., Wallsten, T. S., Erev, I., and Cohen, B. L. (1990) Revision of Opinion with Verbally and Numerically Expressed Uncertainties. *Acta Psychologica*. 74. pp. 61–79.

Reagan, R. T., Mosteller, F., and Youtz, C. (1989) Quantitative Meanings of Verbal Probability Expressions. *Journal of Applied Psychology*. 74. pp. 433–42.

Real, L. A. (1991) Animal Choice Behavior and the Evolution of Cognitive Architecture. *Science*. 253. pp. 980–6.

Ritov, I. and Baron, J. (1990) Reluctance to Vaccinate: Omission Bias and Ambiguity. *Journal of Risk and Uncertainty*. 5. pp. 49–61.

Rode, C., Cosmides, L., Hell, W., and Tooby, J. (1999) When and Why Do People Avoid Unknown Probabilities in Decisions under Uncertainty? Testing Some Predictions from Optimal Foraging Theory. *Cognition*. 72. pp. 269–304.

Russo, J. E. and Schoemaker, P. J. (1992) Managing Overconfidence. *Sloan Management Review*. 33. pp. 7–17.

Schweitzer, M. E. and Hsee, C. K. (2002) Stretching the Truth: Elastic Justification and Motivated Communication of Uncertain Information. *Journal of Risk and Uncertainty*. 25. pp. 185–201.

Seidenfeld, T. Schervish, M. J., and Kadane, J. B. (2011) Forecasting with Imprecise Probabilities. In Coolen, F., de Cooman, G., Fetz, T., and Oberguggenberger, M. (eds.) *Proceedings of the Seventh International Symposium on Imprecise Probability: Theories and Applications*. pp. 317–26. Innsbruck, Austria.

Shafer, G. (1976) *A Mathematical Theory of Evidence.* Princeton, NJ: Princeton University Press.

Simon, H. A. (1956) Rational Choice and the Structure of Environments. *Psychological Review*. 63. pp. 129–38.

Simon, H. (1982) *Models of Bounded Rationality*. 2 vols. Cambridge, MA: MIT Press.

Simpson, R. H. (1944) The Specific Meanings of Certain Terms Indicating Differing Degrees of Frequency. *Quarterly Journal of Speech*. 30. pp. 328–30.

Slovic, P., and Peters, E. (2006) Risk Perception and Affect. *Current Directions in Psychological Science*. 15. 322–5.

Smith, C. A. B. (1961) Consistency in Statistical Inference and Decision. *Journal of the Royal Statistical Society.* B. 23. pp. 1–37.

Smithson, M. (1997) Judgment Under Chaos. *Organizational Behavior and Human Decision Processes*. 69. pp. 59–66.

Smithson, M. (1999) Conflict Aversion: Preference for Ambiguity vs. Conflict in Sources and Evidence. *Organizational Behavior and Human Decision Processes*. 79. pp. 179–98.

Smithson, M. (2013) Conflict and Ambiguity: Preliminary Models and Empirical Tests. In *Proceedings of the Eighth International Symposium on Imprecise Probability: Theories and Applications*. pp. 303–10. Compiègne, France. 2–5 July 2013.

Smithson, M., Budescu, D. V., Broomell, S. B., and Por, H.-H. (2012) Never Say 'Not:' Impact of Negative Wording in Probability Phrases on Imprecise Probability Judgments. *International Journal of Approximate Reasoning*. 53. pp. 1262–70.

Smithson, M. and Segale, C. (2009) Partition Priming in Judgments of Imprecise Probabilities. *Journal of Statistical Theory and Practice*. 3. pp. 169–82.

Smithson, M., Verkuilen, J., Hatori, T., and Gurr, M. (2010) *More than a Mean Difference: New Models and Findings of Partition Priming Effects on Probability Judgments*. Paper presented at the 31st Annual Conference of the Society of Judgment and Decision Making, Nov. 19–22. St Louis, MO.

Soll, J. and Klayman, J. (2004) Overconfidence in Interval Estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 30. pp. 299–314.

Sopena, A. (2009) Somewhere in Between: The Contribution of Ethnicity, Threat and Social Identity to the Production of Marginalizing Racism. Unpublished honours thesis. Canberra: The Australian National University.

Stanovich, K. E. and West, R. F. (2000) Individual Differences in Reasoning: Implications for the Rationality Debate. *Behavioral & Brain Sciences*. 23. pp. 645–65.

Stone, D. R. and Johnson, R. J. (1959) A Study of Words Indicating Frequency. *Journal of Educational Psychology*. 50. pp. 224–7.

Teigen, K. H. (1994) Variants of Subjective Probabilities: Concepts, Norms, and Biases. In Wright, G. and Ayton, P. (eds.) *Subjective Probability*. pp. 211–38. London: Wiley.

Teigen, K. H. and Brun, W. (1995) Yes, but it is Uncertain: Direction and Communicative Intention of Verbal Probabilistic Terms. *Acta Psychologica*. 88. pp. 233–58.

Thüring, M. and Jungermann, H. (1990) The Conjunction Fallacy: Causality vs. Event Probability. *Journal of Behavioral Decision Making*. 3. pp. 61–74.

Tsao, C. J. and Wallsten, T. S. (1994) Effects of the number of outcomes on the interpretation and selection of verbal and numerical probabilities in dyadic decisions. Unpublished manuscript.

Tversky, A. and Kahneman, D. (1974) Judgment under Uncertainty: Heuristics and Biases. *Science*. 185. pp. 1124–31.

Tversky, A. and Kahneman, D. (1983) Extensional versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment. *Psychological Review*. 90. pp. 293–315.

Tversky, A. and Kahneman, D. (1992) Advances in Prospect Theory: Cumulative Representation of Uncertainty. *Journal of Risk and Uncertainty*. 5. pp. 297–323.

Tversky, A. and Koehler, D. (1994) Support Theory: A Nonextensional Representation of Subjective Probability. *Psychological Review*. 101. pp. 547–67.

von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.

Wallsten, T. S., Budescu, D. V., and Erev, I. (1988) Understanding and Using Linguistic Uncertainties. *Acta Psychologica*. 68. pp. 39–52.

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., and Forsyth, B. H. (1986) Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General*. 115. pp. 348–65.

Wallsten, T.S., Budescu, D.V., and Zwick, R. (1993) Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments. *Management Science*. 39. 176–90.

Weber, E. U. and Hilton, D. J. (1990) Contextual Effects in the Interpretations of Probability Words: Perceived Base Rate and Severity of Events. *Journal of Experimental Psychology: Human Perception and Performance*. 16. pp. 781–9.

Weber, E. U., Shafir, S., and Blais, A-R. (2004) Predicting Risk Sensitivity in Humans and Lower Animals: Risk as Variance or Coefficient of Variation. *Psychological Review*. 111. pp. 430–45.

Winman, A., Hansson, P., and Juslin, P. (2004) Subjective Probability Intervals: How to Cure Overconfidence by Interval Evaluation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 30. pp. 1167–75.

Wolford, G., Taylor, H. A., and Beck, J. R. (1990) The Conjunction Fallacy? *Memory and Cognition*. 18. pp. 47–53.

Yaniv, I. and Foster, D. P. (1995) Graininess of Judgment under Uncertainty: An Accuracy-Informativeness Tradeoff. *Journal of Experimental Psychology: General*. 124. pp. 424–32.

Yaniv, I. and Foster, D. P. (1997) Precision and Accuracy of Judgmental Estimation. *Journal of Behavioral Decision Making*. 10. pp. 21–32.

Yates, J. F. and Carlson, B. W. (1986) Conjunction Errors: Evidence for Multiple Judgment Procedures Including 'Signed Summation'. *Organizational Behavior and Human Decision Processes*. 37. pp. 230–53.

Zadeh, L. (1965) Fuzzy Sets. *Information and Control*. 8. pp. 338–53.

CHAPTER 24

PROBABILITY ELICITATION

STEPHEN C. HORA

24.1 The Role of Probability Elicitation

The analysis and solution of complex problems often entails dealing with uncertainty about key factors and values. Sometimes data or first principles are inadequate for characterizing these uncertainties without some interpretation. For example, the data may be from analogues, may be sparse, or may be inconsistent or conflicting in that different studies of phenomena produce varying results. But simple mathematical methods such as averaging values do not take into account the differences in the validities and reliabilities of the competing evidence. In such a situation, the human mind appears to be the best available tool for making sense of the available information. Probability elicitation provides a formal method for quantifying such judgements so that we obtain a comprehensive picture of our current state of knowledge and the inherent uncertainty in that knowledge.

In a study related to mammography, indirect evidence concerning the biological risks of radiation was available to use as a basis for making judgements about relative risks presented by low-energy vs. high-energy radiation. Evidence was drawn from studies of human epidemiology in subjects including bomb survivors and those suffering as a result of occupational risks, studies with animal subjects, laboratory experiments measuring chromosome damage and recombination, and micro-dosimetry. Eight experts were empaneled to construct an uncertainty distribution about the relative damage from various levels of energy and from various sources such as x-rays, cobalt, etc. The ultimate goal was to inform decisions about the risk/reward trade-offs of radiation mammography.

Expert judgement has also played a major role in studies of nuclear power plant safety. An important parameter in models of power plant safety is the pressure at which a containment will fail. But for reasons of practicality and a record with virtually no incidents, experts were employed to interpret design parameters, engineering specifications, and experimental tests on miniature containments, and to bring knowledge of analog systems such as chemical processing vessels to provide probability distributions for the maximum pressure obtainable before failure.

Unfortunately, the formation of probabilistic judgements is a cognitively difficult task and requires preparation. This preparation includes surveying what is known about the issues under consideration and practice in forming probabilistic judgements. The experts that one employs in a probability elicitation then need to be expert in the subject matter (this is sometimes termed substantive expertise) and expert in expressing knowledge in terms of probabilities (normative expertise). Normative expertise does not necessarily come with subject area knowledge and must be developed.

Probability elicitation is not a cure-all nor is it applicable in all situations. It is best used where data are insufficient or require interpretation, where the issues are important enough to warrant careful quantification, and where real expertise exists and can be identified.

24.2 Basic Considerations

Probability elicitation is about representing uncertainty about facts or events. The probabilities and probability distributions reflect both what is known and what is uncertain. It is not about preferences or subjective values. For example, the issue of whether we should reduce the chance of traffic accidents by reducing maximum speed limits requires judgements about the relative desirability of lower accident rates versus shortened travel times. The difficulty is that no definitive resolution is possible. People will agree to disagree. For probability elicitation to be successful, the target event or quantity should be physically measurable, at least conceptually. We should be able to know, eventually, whether an event has occurred or what value a variable has attained. This helps ensure that the question being asked is understood in the same way by both the person asking the question and the expert who provides the answer.

Judgements put in the form of probabilities have the distinct advantage of being in a mathematical form that can be combined with other sources of information and manipulated through the power of mathematics. Judgements given in verbal form not only lack specificity but cannot be coherently integrated into a model. In the study of mammography, for example, numerical uncertainty distributions are essential for crafting policies about the frequency of mammograms and the age at which they should first be taken. Saying that doses of low-energy ionizing radiation are more dangerous than similar doses of high energy, or much more dangerous or somewhat more dangerous for that matter, does not allow one to construct optimal policies that minimize total risk.

The role of expert judgements is to express the state of knowledge and, inevitably, that state of knowledge will change as new evidence becomes available. Thus, these judgements are ephemeral and we should anticipate that they will be refined. They provide a snapshot of what is known at a given point in time.

There are situations where probability elicitation is not appropriate. These include:

1. The absence of evidence upon which the judgements are based
2. An absence of identifiable experts
3. Questions involving events or quantities that are not resolvable, at least conceptually.

24.3 Response Modes

We will distinguish between events and values of variables. Events will resolve into either occurrence or non-occurrence, which can be coded by the indicators X = 1 and X = 0 respectively. An event occurs one-time or not-at-all. When there is a sequence of events, one can measure the relative frequency of X =1 compared to X = 0. The knowledge of the chance of a single event and uncertainty about that event are normally encoded in a single number, the probability. In contrast, a relative frequency is a number in the interval [0,1] and a probability distribution can be assigned to represent the uncertainty about the value the relative frequency will have when resolved.

Probabilities of events can also be encoded as odds – the ratio of the probabilities of occurrence and non-occurrence. Uncertainty about quantities or frequencies can be encoded as densities, probability mass functions, or distribution functions. It is customary to select a response mode that will facilitate the elicitation by better capturing the subject-matter experts' true state of knowledge and at the same time avoiding judgements that are cognitively difficult to make. In some instances, visual devices have been used to provide analogs for probabilities. For example, Stanford Research Institute's decision analysis group pioneered the use of a probability wheel that could be adjusted to show various probabilities and their complements (Merkhofer 1987).

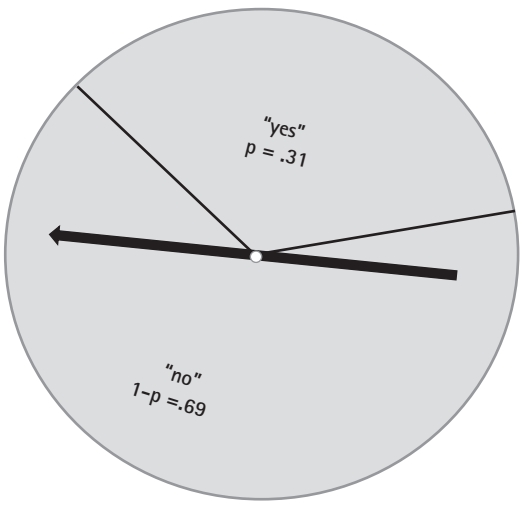


FIGURE 24.1 ProbabilityWheel.

24.4 Selecting and Framing the Issues

For an issue to be a good candidate for probability elicitation, it must meet the requirements of resolvability, a scientific basis from which judgements can be made, and experts,

and it must be of sufficient importance to warrant the resources that are required to engage in a formal elicitation process. When performing an analysis, one most often finds that that a handful of variables or events are responsible for determining the conclusions and effort should be spent on careful quantification of these values, while other variables or events that are less critical can be estimated in a less resource-intensive manner. A primary tool in sifting out those important variables is sensitivity analysis (Saltelli et al. 2004).

How the issues or questions are presented to an expert has the potential to distort their answers and therefore, one should take care not to lead or bias the expert in one direction or another. We will have more to say on this issue in Section 24.8, titled "Biases in Judgement Formation". Clarity is very important in the issue and with complex issues it is sometimes hard to achieve. Both the elicitor and the expert may have unstated assumptions and they may not even think about these assumptions. What is included in the factors that generate uncertainty and what factors are excluded is sometimes the cause of this disconnect. For example, in a study of radioactive deposit after a radiation release from a nuclear power generating station, experts were asked for the rates of wet and dry deposition of the radioactive material on northern European grasslands (Harper et al. 1994). The experts inquired about the length of the grass, as this would be an important factor in estimating the deposition rate. The elicitors had not provided information about this factor and, in fact, the elicitors had assumed that the experts would integrate uncertainty about grass length into uncertainty about the deposition rate.

There are two steps that should be taken to avoid problems that arise because the two parties in an elicitation, the expert and the elicitor, interpret a question differently. The first step is called the clairvoyance test (Spetzler and Stael von Holstein 1975, Stael von Holstein and Matheson 1979) and entails thinking about asking the question of a clairvoyant. If the clairvoyant requires more information before answering the question, then the presentation of the issue is incomplete. Spetzler and Stael von Holstein use the example of asking for the price of wheat and the clairvoyant needing to know the quantity, the type of wheat, the date, at which exchange, and whether the question was about the buying or selling price. The second step is to perform a dry run with stand-in experts. Although this seems like a lot of trouble, poorly presented issues can make a shipwreck out of an otherwise well-thought-out elicitation exercise.

One must also consider how many issues can be addressed by experts. When there are conditions that are made explicit and vary so that probabilities or distributions must be obtained for each combination of conditions, the elicitation agenda can get out of hand. When faced with too many values to quantify, the judgements will become mechanical as the experts look for simple heuristics to get them through all the assessments.

24.5 Selecting Experts

The identification of experts requires that one develop some criteria by which expertise can be measured. Generally, an expert is one who "has or is alleged to have superior knowledge about data, models and rules in a specific area or field" (Bonano et al. 1990). But measuring against this definition requires one to look at indicators of knowledge rather than knowledge *per se*. The following list contains some indicators:

- Research in the area as identified by publications and grants
- Citations of work
- Degrees, awards, or other types of recognition
- Recommendations and nominations from respected bodies and persons
- Positions held
- Membership or appointment to review boards, commissions, etc.

In addition to the above indicators, experts may need to meet some additional requirements. The expert should be free from motivational biases caused by the economic, political, or other interest in the decision. Experts should be willing to participate and they should be accountable for their judgements (Cooke 1991). This means that they should be willing to have their names associated with their specific responses. Many times physical proximity or availability at certain times will be an important consideration.

How the experts are to be organized also impacts the selection. Often, when more than one expert is used, the experts will be redundant of one another – meaning that they will perform the same tasks. In such a case, one should attempt to select experts with differing backgrounds, responsibilities, fields of study, etc., so as to gain a better appreciation of the differences among beliefs. In other instances, the experts will be complementary, each bringing unique expertise to the question. Here, they act more like a team and should be selected to cover the disciplines needed.

Some analyses undergo extreme scrutiny because of the public risks involved. This has been the case with radioactive waste disposal. In such instances, the process for selecting (and excluding) experts should be transparent and well documented. In addition to written criteria, it may be necessary to isolate the project staff from the selection process. This can be accomplished by appointing an independent selection committee to seek nominations and make recommendations to the staff (Trauth, Hora, and Guzowski 1994).

How many experts should be selected? Experience has shown that the differences among experts can be very important in determining the total uncertainty about a question. Clemen and Winkler (1985) examine the impact of dependence among experts using a normal model and conclude that three to five experts are enough. Hora (2004) created synthetic groups from the responses of real experts and found that three to six or seven experts are sufficient with little benefit from additional experts beyond that point. This result is supported by theoretical findings (Hora 2010). When experts are organized in groups and each group provides a single response, then this advice would apply to the number of groups. The optimal number of experts within a group has not been addressed and is certainly dependent on the complexity of the issues being considered.

Another factor that is important in the selection of experts is the freedom from motivational bias. A motivational bias arises from economic or political considerations where the expert's judgement may influence an outcome that the expert desires to control. This leads to the expert to shape their answer in a way that leads to such a desirable outcome. Motivational biases, whether real or potential, can be troublesome, particularly when the most knowledgeable experts are also most subject to such a bias. This situation occurred in the NUREG-1150 study of nuclear power generating stations (Hora and Iman 1989). One of the issues under study was the failure of cooling water pumps. The most knowledgeable expert was also an employee of the pump manufacturer and thus the potential for bias arose. The solution was to allow the potentially biased expert to give testimony to the experts whose judgements were assessed.

24.6 Techniques for Elicitation

The simplest case to begin with is the uncertainty regarding an event that resolves into one of two states – the event occurs or does not occur. The required response is a single probability. It is important to distinguish this situation from a sequence of events where some events in the sequence resolve one way and others resolve another way. With a sequence of events there is a frequency of occurrence that is conceptually knowable and it is proper to create a probability distribution for this frequency. This is not the case for a non-repeating event – one that can occur only once. It is tempting to assign probabilities to probabilities or to use a range for a probability as if the probability had some physical or measurable value even if this is not the case. In any event, even if one was to assign probabilities to probabilities, expectations are linear in probabilities and neither expected values nor expected utilities would differ from those obtained using the mean of the probabilities. For discussions of second-order probabilities and their meaningfulness see de Finetti (1977) and Skyrms (1980). It is perhaps disconcerting that the expert's probability of a non-repeating event simultaneously carries information about an event's chances and its uncertainty. These two aspects of knowledge about a non-repeating event are inseparable, however.

The most straightforward approach to obtaining a probability on a non-repeating event is to ask an expert for that numerical value. An expert who replies with "I don't know the value," is most likely thinking that there is some physical, measurable value that should be known but is not. The probability being sought is a degree of belief and does not have a true, knowable value. It is, instead, a reflection of the expert's knowledge about the chance of the event and will differ from expert to expert and over time as new information is acquired.

Sometimes indirect methods work better. Odds provide a simple re-expression of a probability and the two are easily calculated from each another. Odds require a relative judgement about the chances of an event and its complement rather than a direct judgement resulting in a numerical value. The judgement that an event is, for example, four times more likely to occur than not to occur may be easier for the expert to feel comfortable with than a direct judgement that the event has a .8 probability.

Another type of comparison is the comparison between a physical representation of the probability of an event and a physical representation of the probability of its complement. The aforementioned probability wheel developed for use in probability elicitation by Stanford Research Institute requires such comparisons. This device provides a visual analogue to the probability of an event and its complement. The partial disks can be moved so that one segment is proportional to a number between .5 and 1.0 while the other segment is proportional to the complement. Which segment represents the event and which segment represents the complement is decided by which event is more likely to occur. (Stael von Holstein and Matheson 1979).

Beyond odds and the wheel, analysts have attempted using verbal descriptors of chances such as probable, rare, virtually certain, etc. Kent (1964), head of the CIA's Office of National Estimates, proposed such a verbal scale. But research has shown that these descriptors are interpreted quite differently by various individuals (Druzdel 1989).

Perhaps the best approach to events with multiple outcomes is to decompose the assessment into a number of assessments of events with binary outcomes. Judgements about such events are not as difficult to make. This is a "divide and conquer" approach and will

result in coherent assessments. Decomposition can be accomplished through probability trees, influence diagrams (Shachter 1988), or even formulas. When probability trees are used, the assessed probabilities are conditional probabilities and marginal probabilities of the conditioning events or variables.

There are many decompositions possible for a given problem. One should look for a decomposition that requires judgements that the expert is best prepared to make. Hora, Dodd, and Hora (1993) note that it is possible to over-decompose a problem and make the assessment task more difficult. If time allows, one can look at several decompositions of a given problem and resolve inconsistencies among the recomposed probabilities. Ravinder, Kleinmuntz, and Dyer (1988) examine the propagation of error in subjective probability decompositions.

Many assessments are concerned with the value of a variable. Usually, these variables have a continuous range of potential values. Sometimes the range of a variable is bounded by physical considerations and sometimes it is unbounded. Also, one end of the range may be bounded and the other may be conceptually infinite. For example, the depth of snow at a given location is bounded below by zero but has no well-defined upper limit. Unbounded variables are troublesome in that making judgements about the most extreme possible values is difficult.

In some instances, the decomposition of a probability assessment depends upon the source of uncertainties and the resolvability of those uncertainties. Knight (1921) made the distinction between "risk" (randomness with knowable probabilities) and "uncertainty" (randomness with unknowable probabilities). Today, these components of uncertainty are termed "aleatory" and "epistemic" uncertainties. *Reliability Engineering & System Safety* devoted a special issue to this subject (Helton and Burmaster 1996). Morgan et al. (1992) also discuss various components of uncertainty.

Assessment of continuous distributions can be accomplished through several methods of elicitation, each with its own set of advantages and constraints that depend on the format, scope and intent of an elicitation session. One of the most common methods for eliciting continuous distributions is having the expert provide a number of points on the distribution function (cumulative probability function). We denote such a point by the pair (p,v) where p is the probability that the quantity in question has a value no larger than v, that is $P(X \leq v) = p$ where is X is the uncertain quantity. Of course, one is limited to assessing only a small number of such points, often as few as three or five. The pairs (p,v) may be denoted by v_p and are called fractiles or quantiles. We will use the term fractile rather than quantile, although the two terms are used interchangeably. Fractiles can be assessed either by specifying p and asking for v or by specifying v and asking for p. Both techniques can be used in a probability elicitation, although the first approach has less of a tendency to direct the expert towards specific values, a phenomenon known as the anchoring bias. These two approaches, fixing p and asking for v and vice versa, have been termed the p and v methods (Spetzler and Stael von Holstein 1975).

A variation on the p method is called successive subdivision. It requires the expert to specify a value that breaks an interval into two equally likely subintervals. The process is repeated several times until enough points on the distribution function are obtained to give a good idea of its shape. At the first subdivision, the expert is asked to provide a value such that the uncertain quantity is equally likely to be above as below this value. This assessment yields $v_{.50}$, the median of the distribution. The second stage of successive subdivision entails dividing the range below the median into two equally likely sub-subintervals and

similarly dividing the range above the median into two equally likely sub-subintervals. These subdivisions yield $v_{.25}$ and $v_{.75}$, the first and third quartiles respectively. There are now three values, $v_{.25}$, $v_{.50}$ and $v_{.75}$, that divide the range of possible values into four subintervals each having a probability of .25. Once again some or all of these intervals may be resubdivided until the level of fineness needed to represent the distribution is obtained.

An analyst may choose to use successive subdivision for the median and quartile values and then switch modes to direct assessment using the p or v method for values needed to complete the distribution. Comparisons should also be made to check the consistency of results. For example, the analyst might ask whether it is more or less likely that the quantity is between $v_{.25}$ and $v_{.75}$ than that it is outside that interval. This is a consistency check, and if the analyst finds that inside or outside the interval is more likely, it will be necessary to make some adjustments.

An alternative to the p and v methods is that of selecting a parametric family of distributions and asking the expert to provide judgements about the parameters (such as the mean and standard deviation for a normal density or shape and scale parameters for a gamma density), either directly or indirectly through fractiles or other assessments. Although this approach may be attractive in terms of elicitation duration and simplicity, there are some significant drawbacks to using such an approach. The expert is required to agree to an arbitrary family of distributions prior to elicitation, thus imposing distribution constraints on the judgements. Additionally, the subjective estimates of parameters may require more difficult judgements, which can then lead to greater error being introduced into the assessment process. For example, judgements about a mean may be much more difficult to make than judgements about a median or mode when a distribution is asymmetric. Likewise, judgements about standard deviations may be more difficult to make than judgements about interquartile ranges. To avoid these difficulties, one may ask experts to provide probability judgements on related quantities and then transform these judgements into the intended parameters. For instance, if one was looking to fit a gamma distribution then one might ask an expert for the most likely value (i.e. the mode) and possibly the 0.95 fractile. These values could then be mapped to the distribution, which would be re-shown to the expert to confirm that the new distribution function matches his/her judgement.

24.7 The Quality of Judgements

Because subjective probabilities are personal and vary from individual to individual and from time to time, there is no "true" probability that one might use as a measure of the accuracy of a single elicited probability. For example, consider the question "what is the probability the next elected president of France is a woman?". Individuals may hold different probabilities or degrees of belief about this event occurring. There is, however, no physical, verifiable probability that could be known but remains uncertain. The event will resolve as occurring or not but will not resolve to a frequency or probability.

It is possible to address the goodness of probabilities, however. There are two properties that are desirable to have in probabilities:

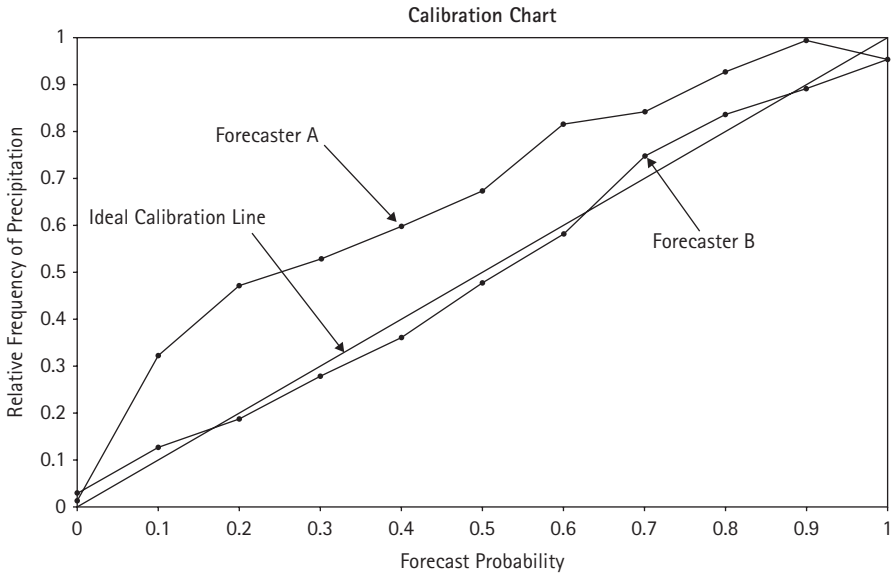


FIGURE 24.2

- Probabilities should be informative
- Probabilities should authentically represent uncertainty

The first property, being informative, means that probabilities closer to 0.0 or 1.0 should be preferred to those closer to .5, as the more extreme probabilities provide greater certainty about the outcome of an event. In a like manner, continuous probability distributions that are narrower or tighter convey more information than those that are less concentrated. The second property, the appropriate representation of uncertainty, requires consideration of a set of assessed probabilities. For those events that are given an assessed probability of p, the relative frequency of occurrence of those events should approach p.

To illustrate this idea, consider two weather forecasters who have provided precipitation forecasts as probabilities. The forecasts are given to a precision of one digit. Thus a forecast of .2 is taken to mean that there is a 20% chance of precipitation. Forecasts from two such forecasters are shown in Figure 24.2 above.

Ideally, each graph would have a 45° line indicating that the assessed probabilities are faithful in that they correctly represent the uncertainty about reality. Weather Forecaster B's graph shows a nearly perfect relation while the graph for Forecaster A shows poorer correspondence between the assessed probabilities and relative frequencies, with the actual frequency of rain exceeding the forecast probability. The graph is not even monotonic at the upper end.

Graphs showing the relation between assessed probabilities and relative frequencies are called calibration graphs and the quality of the relationship, loosely called calibration, can be good or poor (Lichtenstein, Fischhoff, and Phillips 1982). Calibration graphs can also be constructed for continuous assessed distributions. Following Hora (2004), let $F_i(x)$ be a set of assessed continuous probability distribution functions and let x_i be the corresponding actual values of the variables. If an expert is perfectly calibrated, the cumulative probabilities of the actual values measured on each corresponding distributions function, $p_i = F_i(x)$, will

be uniformly distributed on the interval [0,1] and thus the empirical distribution function will plot as a 45° line.

Although calibration is an important property of a set of probabilities or probability distributions, it is not sufficient, as the probabilities or probability distributions may not be informative. For example, in an area where it rains on 25% of days, a forecaster who always predicts a 25% chance of rain will be perfectly calibrated but provide no information from day to day about the relative chance of rain. But information and calibration are somewhat at odds. Increasing the information by making probabilities closer to zero or one or by making distributions tighter may reduce the level of calibration.

One approach to measuring the goodness of probabilities is through scoring rules. Scoring rules are functions of the assessed probabilities and the true outcome of the event or value of the variable that measure the goodness of the assessed distribution and incorporate both calibration and information into the score. The term "strictly proper scoring function" refers to the property that the expected value of the function is maximized when the probabilities or probability functions to which the function is applied are identical to the probabilities or probability functions that are used to take the expectation. An example will clarify.

A simple strictly proper scoring rule for the assessed probability of an event is the Brier or quadratic rule (Brier 1950):

$$\begin{aligned} S(p) &= -(1-p)^2 \text{ if the event occurs} \\ &= -p^2 \text{ if the complement of the event occurs.} \end{aligned}$$

where p is the assessed probability. For any probability q, the mathematical expectation $E_q[S(p)] = -q(1-p)^2 - (1-q)p^2$ is maximized with respect to p by setting $p = q$. Thus, if an expert believes the probability is q, that expert will maximize the perceived expectation by responding with q. In contrast, the scoring rule $S(p) = -p$ if the event occurs and $S(p) = -(1-p)$ if it does not occur, while intuitively pleasing, does not promote truthfulness. Instead, the expected score is maximized by providing a probability p of either 0.0 or 1.0, depending on whether q is less than or larger than 0.5. Winkler (1996) provides a discussion of this Brier rule and other strictly proper scoring rules.

The concept of a strictly proper scoring rule can be extended to continuous distributions (Matheson and Winkler 1976). For example, the counterpart of the quadratic scoring rule for continuous densities is:

$$S[f(x), w] = 2f(w) - \int_{-\infty}^{\infty} f^2(x)dx$$

Expected scores can be decomposed into recognizable components. The quadratic rule for continuous densities can be decomposed in the following manner. Suppose that an expert's uncertainty is correctly expressed through the density $g(x)$ but the expert responds with $f(x)$, either through inadvertence or intention. The expected score can be written as:

$$\begin{aligned} &E_g\{S[f(x), w]\} = I(f) - C(f,g) \\ &\text{where } I(f) = \int_{-\infty}^{\infty} f^2(x)dx \text{ and } C(f,g) = 2\int_{-\infty}^{\infty} f(x)[f(x) - g(x)]dx \end{aligned}$$

I(f) is the expected density associated with the assessed distribution and is a measure of information. *C(f,g)* is a non-negative function that increases as *g(x)* diverges from *f(x)*. Thus *C(f,g)* is a measure of miscalibration. Further discussion of decomposition can be found in Murphy (1972, 1973). Haim (1982) provides a theorem that shows how a strictly proper scoring rule can be generated from a convex function. See also Savage (1971).

24.8 Biases in Judgement Formation

The process of expressing one's knowledge in terms of probabilities is not simple and has been shown to be subject to some types of repeatable errors. Cognitive psychologists have detected, classified, and analyzed these errors in experimental settings. This work dates back to the 1970s and was spearheaded by Kahneman and Tversky (Bar-Hillel 2001). Kahneman, Slovic, and Tversky (1982) provide a compilation of research in this area up to 1982 while Gilovich, Griffin, and Kahneman (2002) contains contributions from the next two decades.

Judgemental errors are thought to be related to the way information is processed, that is, the heuristics used in forming the judgements. Two predominant heuristics have been labeled "the representiveness heuristic" and "the availability heuristic".

Representativeness is the process of using some relevant cues to associate a target event or quantity with a similar set of targets. Bar-Hillel (2001) notes, however, that similarity judgements obey different rules to those governing probability judgements. For example, subjects given a description of what could possibly be a location in Switzerland respond with a higher probability that the location is in Switzerland than the probability that the location is in Europe (Bar-Hillel and Neter 1993). Such a judgement is irrational as one event is entirely included within another (being in Switzerland implies being in Europe.)

Another manifestation of the representativeness heuristic is termed "the base rate bias". For example, if a diagnostic test for a disease has error rates of 2% (both false positive and false negative) and a person has a positive result, they may wrongly conclude that there is a 98% chance they have the disease. The actual probability depends heavily on the base rate of the disease in the population and it may be that most positive results are actually false positives. Other effects related to the representativeness heuristic include failure to regress, failure to consider sample size, and incorrect interpretations of randomness (O'Hagan et al. 2006).

A second major class of biases arises from the availability heuristic. "Availability" denotes the ability to access or recall information. Cues that are easier to recall tend to be given more weight in forming probability judgements. An experiment that illustrates this effect entails showing a subject a list of names of celebrities. Manipulating the list so that the members of one sex are somewhat more famous than the members of the opposite sex will result in the subject's overestimating the relative frequency of names on the list belonging to the sex having the more famous members (Tversky and Kahneman 1973). Another manifestation of availability occurs when subjects are asked to estimate the relative frequencies of various causes of death. Subjects tend to overestimate the frequencies of sensational causes and underestimate the frequencies of mundane causes. For instance, the chance of death from a stroke will be underestimated, while the chance of death by firearms will be overestimated.

Information about death by firearms is more widely publicized – every such death will receive media attention, while death from stroke is likely to be reported only in the case of well-known individuals. Thus, the mind finds it easier to recall instances of death by firearms or, just as importantly, overestimates the frequency by overestimating the number of such instances that could be brought to mind if one really tried (Fischhoff and MacGregor 1982).

Anchoring, which is particularly salient in probability elicitation, occurs when a subject gives too much credit to a reference point, so that other possible references, or the fallibility of the selected reference, are ignored (Tversky and Kahneman 1974). For example, two experts who are equally qualified and have conducted studies to determine a particular quantity are likely each to give more credit to their own work and less credit than is warranted to their colleague's work. Anchoring can also occur when there is sparse evidence. The expert may rely too heavily on evidence that is available and ignore how new evidence could change our judgements.

Consider a situation where there are two known sources of information, each leading to a somewhat different estimate of a quantity, and there is a third source of information not known to the expert at this time. Imagine the three pieces of evidence are represented by dots on a straight line. Suppose that one of the dots is initially invisible. If the invisible dot is chosen randomly from the three dots, it is twice as likely that the invisible dot is outside the range of the visible dots. Thus, the appearance of the a third piece of information is twice as likely to spread the range of possibilities as to confirm the range if one "anchors" the range using the observable values. The paradox is that in judgement formation, information may result in an increased expression of uncertainty, so that the more we know, the less we say we know.

Anchoring can also occur as the result of a scale or instrument that suggests values for a variable. Subjects are uncomfortable going outside the set of suggested values. Moreover, the subject is drawn to the center of the scale as a representative or central value. Hora, Hora, and Dodd (1992) and Winman, Hansson, and Juslin (2004) have shown that asking for probabilities of fixed intervals rather than asking for intervals with fixed probabilities can result in anchoring.

Another scale effect can occur when there are discrete, mutually exclusive outcomes and one of the outcomes is a catchall, "everything else" category. For example, in a hospital study, a description of an admittee's symptoms was given along with a list of possible diagnoses. Two lists were used, one list with four named diagnoses and a catchall category to include all unnamed diagnoses. The second list had three possibilities, two of the four diagnoses named in the first list and a catchall category that implicitly contained the missing two diagnoses from the first list. The result is that subjects given the second list gave a lower probability for the catchall category than the sum of the probabilities given to the two missing diagnoses and the catchall category on the first list. These probabilities should be equivalent. The availability of the additional diagnoses has been attributed as the reason for this incongruity. This has been termed the packing effect. One of the first studies in this area involved an experiment using automobile fault trees (Fischhoff, Slovic, and Lichtenstein 1978). In this study some potential causes of automobile failure were omitted from a diagram rather than being relegated to a catchall category. The subjects were asked to estimate the total probability of the missing causes and consistently underestimated this probability. In this study, the bias was termed the pruning effect. Fox and Clemen (2005) provide an alternative explanation of this bias.

The psychological bias that has received the most attention in probability elicitation is the tendency to assign probabilities that are too close to zero or one or to give uncertainty distributions that are too narrow. This bias is called "apparent overconfidence" as the effect is to provide answers that are apparently more certain than is warranted. Some research in this area has employed probabilities for binary events based on almanac data or sensory experiments. For example, a subject might be asked to say which mountain is the higher, Mt. Fuji or the Matterhorn, and then assign a probability to the subject's answer being correct (Lichtenstein and Fischhoff 1977). Keren (1988) conducts a similar experiment but asks the subjects to identify letters while manipulating the difficulty of the task. While subjects tend to assign probabilities that are too high, the effect is much more pronounced when the questions are, in some sense, difficult rather than easy.

There have been other studies that have examined elicited distributions for continuous quantities. Lichtenstein, Fischhoff, and Phillips (1982) provide a table of the findings of such studies by concentrating on the faithfulness of the tail probabilities. In each of the studies, some extreme fractiles, such as the .01 and .99 fractiles, were assessed and in most of the studies the interquartile range was assessed. The studies employed almanac data with known answers or values that would be known shortly, such as football scores (Winkler 1971). Almost uniformly, each study reported that the number of times the target values occurred in these extreme tails was greater than would be indicated by the probabilities – usually much greater. For example, in a study of stock price prediction by graduate business students, Stael von Holstein (1972) reports that 30% of the responses fell above the .99 fractile or below the .01 fractile while 27% fell in the interquartile range rather than the desired 50%. Klayman et al. (1999) find that overconfidence is more pronounced in assessment of continuous quantities (interval assessments) than the assessment of simple questions such as two-choice questions.

Overconfidence can be very troubling where a correct expression of uncertainty is needed. In such circumstances it is wise to educate the expert about this bias and give some training and feedback on the expert's performance. While not every individual will exhibit this bias, it appears to exist in every group of subjects, whether they be experts or students. There is ample evidence (Alpert and Raiffa 1982) to show that probability exercises and feedback will reduce this bias, although not entirely eliminate it.

Overconfidence is associated with the hard-easy effect wherein subjects responding to more difficult questions exhibit more overconfidence than those responding to easy questions (Lichtenstein and Fischhoff 1977). Apparent overconfidence observed in almanac tests has been explained as a manifestation of the hard-easy effect by Juslin (1994). See also Gigerenzer (1991) and Gigerenzer, Hoffrage, and Kleinbolting (1991). A review is provided in Brenner, Koehler, and Liberman (1996).

Another active area of research is support theory, which was introduced by Tversky and Koehler (1994/2002) to explain why elicited probabilities depend on the manner of presentation of the issue and apparently do not conform to the normative principle of additivity of probabilities across mutually exclusive events. The theory envisions probability judgement formation as consisting of three sets of elements: mutually exclusive hypotheses, evidence that supports the hypotheses, and expressed probabilities. The hypotheses are descriptions of events rather than the events themselves. The evidence or support is the perceived strength of the evidence for a particular hypothesis. The judged probability is the weight in favor of a particular hypothesis relative to the weights of all hypotheses.

The weight of evidence, or support for a hypothesis A is given by $s(A)$. The real content of support theory is in the assumption that support is subadditive. Mathematically, if A_1 and A_2 are a partition of A, that is $A = A_1 \cup A_2$ and $A_1 \cap A_2$ is empty, then support theory requires that $s(A) \leq s(A_1) + s(A_2)$. If the inequality holds strictly, then one will find the judged probability of an event to be less than the sum of the probabilities of the constituent elements of the partition of that event. Of course, this is relevant to probability elicitation, as the way the question is posed may influence the response given. To make things more concrete, let A be homicide, A_1 be homicide by an acquaintance, and A_2 be homicide by a stranger. Subadditivity operates if the probability of A is less than the sum of the probabilities of A_1 and A_2. This has been shown to be the case in numerous studies (Tversky and Koehler 1994/2002, Fox and Tversky 1998).

24.9 Using Multiple Experts

It is often beneficial to use more than one expert. The differences in the judgements provided by multiple experts provides a better perspective on the true uncertainty about an event or quantity. This is because the experts have somewhat different knowledge bases and, as discussed in the section on biases, experts tend to underestimate uncertainty, appearing more confident than they should. Thus, aggregated probabilities or distributions may provide a better representation of our knowledge. Another advantage of combining judgements is that it provides a consolidated view that can be used in an analysis rather than multiple results corresponding to the judgements of the individuals.

There are two classes of aggregation methods, behavioral and mathematical. Behavioral approaches entail negotiation to reach a representative or consensus distribution. Mathematical methods, in contrast, are based on a rule or formula. The approaches are not entirely exclusive however, as they may both be used to greater or lesser degree to perform an aggregation.

Perhaps the best known behavorial approach is the "Delphi" technique, developed at the Rand Corporation in the 1960s by Dalkey (1967). In the Delphi method, the interaction among the experts is tightly controlled. In fact, they do not meet face to face but remain anonymous to one another. This is done to eliminate the influence that an individual might have because of position or personality. Judgements are exchanged among the experts along with the reasoning for their judgements. After viewing all judgements and rationales, the experts are given the opportunity to modify their judgements. The process is repeated – exchanging judgements and revising them – until the judgements become static or have converged to a consensus. It will often be necessary to apply some mathematical rule to complete the aggregation process.

Another behavioral approach is called the "nominal group" technique (Delbecq, van de Ven, and Gustafson 1986). This technique, like Delphi, controls the interaction among experts. The experts meet and record their ideas or judgements in written form without discussion. A moderator then asks each expert to provide their idea or judgement and records this on a public media display such as a white board, flipchart, or computer screen. There may be several rounds of ideas/judgements, which are then followed by discussion. The judgements or ideas are then ranked individually and anonymously by the experts and the moderator summarizes the rankings. This process can be followed by further discussion and re-ranking or voting on alternatives.

Kaplan (1990) proposes a behavioral method for combining judgements through negotiation with a facilitator. The facilitator and experts meet to discuss the problem. The facilitator's role is to bring out information from the experts and interpret a "consensus body of evidence" that represents the aggregated wisdom of the group.

A wide range of mathematical methods for combining probability judgements has been proposed. Perhaps the simplest and most widely used is a simple average, termed the "linear opinion pool" (Stone 1961). This technique applies equally well to event probabilities and continuous probability densities or distribution functions. It is important to note that with continuous distributions, it is the probabilities, not the values that are averaged. For example, it is tempting, given several medians, to average the medians; but this is not the approach we are referring to. An alternative to the simple average is to provide differential weights to the various experts, ensuring that the weights are non-negative and sum to one. The values of the weights may be assigned by the staff performing the combination of experts or they may result from some measure of the experts' performance. Cooke (1991) suggests that evidence of the quality of probability assessments be obtained using training quizzes with questions from the subject area the expert is addressing. The experts are given weights based on the product of the p-value for the χ^2 test of calibration and the information as measured by the entropy in the assessments. A cut-off value is used so that poorly calibrated experts are not included in the combination.

The most elegant approach to combining judgements is provided by Morris (1977). In his approach, a decision maker assigns a joint likelihood function to the various responses that a group of experts might provide and a prior distribution for the quantity or event in question. The likelihood function is conditional on the quantity or event of interest. Also see French (1985) for a discussion of the axiomatic approaches to combining experts using a Bayesian approach. The decision maker can then develop the posterior distribution for the uncertain quantity or event using Bayes' theorem.

Various mathematical methods for combining probability judgements have different desirable and undesirable properties. Genest and Zidek (1986) describe the following property:

> *Strong set-wise function property*: A rule for combining distributions has this property if the rule is a function only of the assessed probabilities and maps $[0,1]^n \rightarrow [0,1]$. In particular, the combination rule is not a function of the event or quantity in question.

This property in turn implies the following two properties:

> *Zero set property*: If each assessor, $i = 1,\ldots n$ provides $P_i(A) = 0$, then the combined result, $P_c(A)$, should also concur with $P_c(A) = 0$.

> *Marginalization property*: If joint probabilities are combined, the marginal probabilities from the combined distribution will be the same as the combined marginal probabilities.

The strong set property also implies that the combining rule is a linear opinion pool or weighted average of the form

$$P_c(A) = \sum_{i=1}^{n} \alpha_i P_i(A)$$

where the weights, α_i, are non-negative and sum to one.

Another property is termed the independence property:

$$\textit{The independence property: } P_c(A \cap B) = P_c(A)P_c(B)$$

whenever all experts assess A and B as independent events.

But the linear opinion pool does not have this property. Moreover, the linear opinion pool given above cannot be applied successfully to both joint probabilities and the component marginal and conditional probabilities. That is,

$$P_c(A|B)P_c(B) \neq \sum \alpha_i P_i(A|B)P_i(B)$$

except when one of the weights is one and all others are zero, so that one expert is a "dictator".

The strong set property was used by Dalkey (1972) to prove an impossibility theorem for combining rules. Dalkey adopts seven assumptions that lead to the conclusion that "there is no aggregation function for individual probability estimates which itself is a probability function". Boardley and Wolff (1981) argue that one of these assumptions, the strong set property, is unreasonable and should not be used as an assumption.

While the linear rule does not conform to the independence property, its cousin, the geometric or logarithmic rule, does. This rule is linear in the log probabilities and is given by

$$P_c(A) = k\prod_{i=1}^{n} P_i(A)^{\alpha_i} \text{ where } \alpha_i > 0, \sum_{i=1}^{n} \alpha_i = 1$$

and k is a normalizing constant.

The geometric rule also has the property of being externally Bayesian.

Externally Bayesian: The result of performing Bayes' Theorem on individual assessments and then combining the revised probabilities is the same as combining the probabilities and then applying Bayes' Theorem.

While the geometric rule is externally Bayesian, it is also dictatorial in the sense that if one expert assigns $P_i(A) = o$, the combined result is necessarily $P_c(A) = o$. We note that the linear opinion pool is not externally Bayesian.

It is apparent that all the desirable mathematical properties of combining rules cannot be satisfied by a single rule. The topic of selecting a combining method remains open for investigation. However, in practice the most commonly used method is to average probabilities of events or to average densities or cumulative probabilities at various variable values. Although this method works well, it is known to provide aggregated probability density functions that are too widely spread when the experts are well calibrated (Hora 2004, Ranjan and Gneiting 2010). Hora et al. (2013) suggest the median aggregation instead of averaging as this reduces the spreading effect of aggregation.

Empirical studies of aggregation have been most often conducted using forecasts rather than probabilities (Clemen and Winkler 2007). In those situations where probabilities or probability distributions were aggregated and the performance of the aggregate measured against realized values, the results are mixed. Some studies found behavioral methods to perform better while others found mathematical aggregation best. However, group probabilities arrived at by some consensus process do not appear to perform better than

those derived by simple averaging of probabilities. A common, but not universal, finding in these studies is that simpler methods either perform as well as or better than their more complicated counterparts. Decomposition of complex issues into simpler issues has also been found to improve the quality of the assessed probabilities.

24.10 Organizing Elicitations

In addition to defining issues and selecting and training the expert(s), it is necessary to address a number of questions concerning the structure for a probability elicitation. These include:

- the amount of interaction and exchange of information among experts
- the type and amount of preliminary information to be provided to the experts
- the time and resources that will be allocated to preparation of responses
- venue – the experts' places of work, the project's home, or elsewhere
- Will there be training, what kind, and how will it be accomplished?
- Are the names of the experts to be associated with their judgements and will individual judgements be preserved and made available?

Decisions about these choices result in the creation of a design for elicitation that has been termed a protocol. Some protocols are discussed in Morgan et al. (1992), Merkhofer (1987), Keeney and von Winterfeldt (1991), and Cooke (1991). We will briefly outline two different protocols that illustrate the range of options that have been employed in expert elicitation studies.

Morgan et al. (1992) identify the Stanford Research Institute (SRI) assessment protocol as, historically, the most influential in shaping structured probability elicitation. This protocol is summarized in Spetzler and Stael von Holstein (1975). It is designed around a single expert (subject) and single analyst engaged in a five-stage process. The stages are:

- Motivating – Rapport with the subject is established and possible motivational biases explored.
- Structuring – The structure of the uncertainty is defined.
- Conditioning – The subject is conditioned to think fundamentally about his judgement and to avoid cognitive biases.
- Encoding – This is the actual quantification in probabilistic terms.
- Verifying – The responses obtained in the encoding are checked for consistency.

The role of the analyst in the SRI protocol is primarily to help the expert avoid psychological biases. The encoding of probabilities roughly follows a script. Stael von Holstein and Matheson (1979) provide an example of how an elicitation session might go forward.

A distinguishing feature of the SRI Protocol is the use of the probability wheel described earlier. The encoding stage for continuous variables is described in some detail in Spetzler and Stael von Holstein (1975). It begins with assessment of the extreme values of the values of the variable. An interesting sidelight is that after assessing these values, the subject is

asked to describe scenarios that might result in values of the variable outside the interval and to provide a probability of being outside the interval. The process next goes to a set of intermediate values whose cumulative probabilities are assessed with the help of the probability wheel. Then an interval technique is used to obtain the median and quartiles. Finally, the judgements are verified by testing for coherence and conformance with the expert's beliefs.

While the SRI protocol was designed for solitary experts, a protocol developed by Sandia Laboratories for the U.S. Nuclear Regulatory Commission (Hora and Iman 1989, Ortiz et al. 1991) was designed to bring multiple experts together. The Sandia protocol consists of two meetings.

- First meeting
    - presentation of the issues and background materials
    - discussion by the experts of the issues and feedback on the questions
    - a training session including feedback on judgements.

The first meeting is followed by a period of individual study of approximately one month.

- Second meeting
    - discussion by the experts of the methods, models, and data sources used
    - individual elicitation of the experts.

The second meeting is followed by documentation of rationales and opportunity for the experts to give feedback. The final individual judgements are then combined using simple averaging to the final probabilities or distribution functions.

There are a number of significant differences between the SRI and Sandia protocols. First, the SRI protocol is designed for isolated experts, while the Sandia protocol brings multiple experts together and allows them to exchange information and viewpoints. They are not allowed, however, to view or participate in the individual encoding sessions, nor comment on one another's judgements. Secondly, in the SRI protocol, it is assumed that the expert is fully prepared in that no additional study, data acquisition, or investigation is needed. Moreover, the SRI protocol gives the analyst the task of identifying biases and assisting the expert in counteracting these biases, while the Sandia protocol employs a structured training session to help deal with these issues. In both protocols, the encoding is essentially the same, although the probability wheel is today seldom employed by analysts. Thirdly, the Sandia protocol places emphasis on obtaining and documenting multiple viewpoints, which is consistent with the public policy issues addressed in those studies to which it has been applied.

References

Alpert, M. and Raiffa, H. (1982) A progress report on the training of probability assessors. In Kahneman, D., Slovic, P., and Tversky, A. (eds.) *Judgment under Uncertainty: Heuristics and Biases*. PP. 294–305. Cambridge: Cambridge University Press.

Bar-Hillel, M. (2001) Subjective probability judgments. In Smelser, N. J. and Baltes, D. B. (eds.) *International Encyclopedia of the Social & Behavioral Sciences*. pp. 15248–51. Amsterdam: Elsevier Science Ltd.

Bar-Hillel, M. and Neter, E. (1993) How alike is it versus how likely is it: A disjunction fallacy in stereotype judgments. *Journal of Personality and Social Psychology*. 65. pp. 1119–32.

Boardley, R. F. and Wolff, R. W. (1981) On the aggregation of individual probability estimates. *Management Science*. 27. pp. 959–64.

Bonano, E. J., Hora S. C., Keeney, R. L., and von Winterfeldt, D. (1990) *Elicitation and Use of Expert Judgment in Performance Assessment for High-Level Radioactive Waste Repositories, NUREG/CR-5411*. Washington, DC: U.S. Nuclear Regulatory Commission.

Brenner, L. A., Koehler, D. J., and Liberman, V. (1996) Overconfidence in probability and frequency judgments: A critical examination. *Organizational Behavior and Human Decision Processes*. 65. pp. 212–19.

Brier, G. (1950) Verification of weather forecasts expressed in terms of probabilities. *Monthly Weather Review*. 76. pp. 1–3.

Clemen, R. T. and Winkler R. L. (1985) Limits for the precision and value of information from dependent sources. *Operations Research*. 33. pp. 427–42.

Clemen, R. T. and Winkler, R. L. (2007) Aggregating probability distributions. In *Advances in Decision Analysis: From Foundations to Applications*. pp. 154–76. Cambridge: Cambridge University Press.

Cooke, R. M. (1991) *Experts in Uncertainty*. Oxford: Oxford University Press.

Dalkey, N. (1967) *Delphi*. Rand Corporation Report. Santa Monica, CA: The Rand Corporation.

Dalkey, N. (1972) *An Impossibility Theorem for Group Probability Functions*. Santa Monica, CA: The Rand Corporation.

de Finetti, B. (1977). Probabilities of probabilities: A real problem or a misunderstanding? In Aykac, A. and Brumat, C. (eds.) *New Developments in the Application of Bayesian Methods*. Amsterdam: North Holland.

Delbecq, A. L., Van de Ven, A. H., and Gustafson, D. H. (1986) *Group Techniques for Program Planning: A Guide to Nominal Group and Delphi Processes*. Middleton, WI: Green Briar Press.

Druzdel, M. J. (1989) *Verbal Uncertainty Expressions: Literature Review*. Technical Report CMU-EPP-1990-02-02. Pittsburgh, PA: Department of Engineering, Carnegie Mellon University.

Fischhoff, B. and MacGregor, D. (1982) Subjective confidence in forecasts. *Journal of Forecasting*. 1. pp. 155–72.

Fischhoff, B., Slovic, P., and Lichtenstein, S. (1978) Fault trees: Sensibility of estimated failure probabilities to problem representation. *Journal of Experimental Psychology: Human Perception and Performance*. 4. pp. 330–44.

Fox, C. R. and Clemen, R. T. (2005) Subjective probability assessment in decision analysis: Partition dependence and bias toward the ignorance prior. *Management Science*. 51. 9. pp. 1417–32.

Fox, C. R. and Tversky, A. (1998) A belief-based account of decision under uncertainty. *Management Science*. 44. pp. 879–95

French, S. (1985) Group consensus probability distributions: A critical survey. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M. (eds.) *Bayesian Statistics 2*. pp. 183–201. Amsterdam: North-Holland:.

Genest, C. and Zidek, J. V. (1986) Combining probability distributions: A critique and annotated bibliography. *Statistical Science*. 1. pp. 114–48.

Gigerenzer, G. (1991) How to make cognitive illusions disappear: Beyond heuristics and biases. In Stroebe, W. and Hewstone, M. (eds.), *European Review of Social Psychology*. Vol. 2. New York, NY: Wiley.

Gigerenzer, G., Hoffrage, U., and Kleinbolting, H. (1991) Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*. 98. pp. 506–28.

Gilovich, T., Griffin, D., and Kahneman, D. (2002) *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge: Cambridge University Press.

Haim, E. (1982) *Characterization and Construction of Proper Scoring Rules.* Unpublished doctoral dissertation, University of California, Berkeley.

Harper, F. T., Hora, S. C., Young, M. L., Miller, L. A., Lui, C. H., McKay, M. D., Helton, J. C., Goossens, L. H. J., Cooke, R. M., Pasler-Sauer, J., Kraan, B., and Jones, J. A. (1994) *Probability Accident Consequence Uncertainty Analysis.* Vols. 1–3. (*NUREG/ CR-6244, EUR 15855 EN*). Brussels: USNRC and CEC DG XII.

Helton, J. C. and Burmaster, D. E. (eds.) (1996) Special issue on treatment of aleatory and epistemic uncertainty. *Reliability Engineering & System Safety*. 54. 2–3.

Hora, S. C. (2004) Probability judgments for continuous quantities: Linear combinations and calibration. *Management Science*. 50. pp. 597–604.

Hora, S. C. (2010) An analytic method for evaluating the performance of aggregation rules for probability densities. *Operations Research*. 58. 5. pp. 1440–9.

Hora, S., Dodd, N. G., and Hora, J. (1993) The use of decomposition in probability assessments of continuous variables. *The Journal of Behavioral Decision Making*. 6. pp. 133–47.

Hora, S. C., Fransen, B. R., Hawkins, N., and Susel, I. (2013) Median aggregation of distribution functions. *Decision Analysis*. 10. 4. pp. 279–91.

Hora, S. C., Hora, J. A., and Dodd, N. G. (1992) Assessment of probability distributions for continuous random variables: A comparison of the bisection and fixed value methods. *Organizational Behavior and Human Decision Processes*. 51. pp. 133–55.

Hora, S. C. and Iman, R. L. (1989) Expert opinion in risk analysis: The NUREG-1150 experience. *Nuclear Science and Engineering*. 102. 4. pp. 323–31.

Juslin, P. (1994) The overconfidence phenomenon as a consequence of informal experimenter guided selection of almanac items. *Organizational Behavior and Human Decision Processes*. 57. pp. 226–46.

Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Kaplan, S. (1990) 'Expert Information' vs 'Expert Opinions': Another approach to the problem of eliciting/combining/using expert knowledge in PRA. *Journal of Reliability Engineering and System Safety*. 39. pp. 61–72.

Keeney, R. and von Winterfeldt, D. (1991) Eliciting probabilities from experts in complex technical problems. *IEEE Transactions on Engineering Management*. 38. pp. 191–201.

Kent, S. (1994) Words of estimated probability. In Steury, D. P. (ed.) *Sherman Kent and the Board of National Estimates: Collected Essays*. Washington, DC: Center for the Study of Intelligence.

Keren, G. (1988) On the ability of monitoring non-veridical perceptions and uncertainty knowledge: Some calibration studies. *Acta Psychologica*. 67. pp. 95–119.

Klayman, J., Soll, J. B., Gonzalez-Vallejo, C., and Barlas, S. (1999) Overconfidence: It depends on how, whom you ask. *Organizational Behavior and Human Decision Processes*. 79. pp. 216–47.

Knight, F. H. (1921) *Risk, Uncertainty, and Profit*. Boston, MA: Houghton Mifflin Company.

Lichtenstein, S. and Fischhoff, B. (1977) Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance*. 20. 159–83.

Lichtenstein, S., Fischhoff, B., and Phillips, L. D. (1982) Calibration of probabilities: The state of the art to 1980. In Kahneman, D., Slovic, P., and Tversky, A. (eds.) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.

Matheson, J. E. and Winkler, R. L. (1976) Scoring rules for continuous probability distributions. *Management Science*. 22. pp. 1087–96.

Merkhofer, M. W. (1987) Quantifying judgmental uncertainty: Methodology, experiences, and insights. *IEEE Transactions on Systems, Man, and Cybernetics* 17. pp. 741–52.

Morgan, M. G., Henrion, M., and Small, M. (1992) *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Paperback ed. New York: Cambridge University Press. [Latest printing (with revised Chapter 10) 1998.]

Morris, P. A. (1977) Combining expert judgments: A Bayesian approach. *Management Science*. 23. 7. pp. 679–93.

Murphy, A. H. (1972) Scalar and vector partitions of the probability score. Part I: Two-state situation. *Journal of Applied Meteorology*. 11. pp. 273–82.

Murphy, A. H. (1973) A new vector partition of the probability score. *Journal of Applied Meteorology*. 12. 595–6.

O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., Oakley, J. E., and Rakow, T. (2006) *Uncertain Judgements: Eliciting Experts' Probabilities*. New York, NY: John Wiley.

Ortiz, N. R., Wheeler, T. A., Breeding, R. J., Hora, S., Meyer, M. A., and Keeney, R. L. (1991) The use of expert judgment in the NUREG-1150. *Nuclear Engineering and Design*. 126. pp. 313–31.

Ranjan, R. and Gneiting, T. (2010) Combining probability forecasts. *Journal of the Royal Statistical Society*. B (Statistical Methodology). 72. 1. pp. 71–91.

Ravinder, H. V., Kleinmuntz, D. N., and Dyer, J. S. (1988) The reliability of subjective probability assessments obtained through decomposition. *Management Science*. 34. pp. 186–99.

Saltelli, A., Tarantola, S., Campolongo, F., and Ratto, M. (2004) *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. Chichester: John Wiley.

Savage, L. J. (1971) The elicitation of personal probabilities and expectations. *Journal of the American Statistical Association*. 66. pp. 783—801.

Shachter, R. D. (1988). Probabilistic inference and influence diagrams. *Operations Research* 36. 4. pp. 589–604.

Skyrms, B. (1980) Higher order degrees of belief. In Mellor, D. H. (ed.) *Prospects for Pragmatism: Essays in Honor of F. P. Ramsey*. pp. 109–37. Cambridge: Cambridge University Press.

Spetzler, C. S. and Stael von Holstein, C-A. S. (1975) Probability encoding in decision analysis. *Management Science*. 22. pp. 340–58.

Stael von Holstein, C-A. S. (1972) Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*. 8. pp. 139–58.

Stael von Holstein, C-A. S. and Matheson, J. E. (1979) *A Manual for Encoding Probability Distributions*. Menlo Park, CA: SRI International.

Stone, M. (1961) The linear opinion pool. *Annals of Mathematical Statistics*. 32. pp. 1339–42.

Trauth, K. M., Hora, S. C., and Guzowski, R. V. (1994) A Formal Expert Judgment Procedure for Performance Assessments of the Waste Isolation Pilot Plant, SAND93-2450. Albuquerque, NM: Sandia National Laboratories.

Tversky, A. and Kahneman, D. (1973) Availability: A heuristic for judging probability. *Cognitive Psychology*. 5. pp. 207–32. (Also (in an abbreviated form) in Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.)

Tversky, A. and Kahneman, D. (1974) Judgment under uncertainty: Heuristics and biases. *Science*. 185. 4157. pp. 1124–31. (Also in Kahneman, D., Slovic, P., and Tversky, A. (eds.) (1982) *Judgment under Uncertainty: Heuristics and Biases*. Cambridge: Cambridge University Press.)

Tversky, A. and Koehler, D. J. (1994/2002) Support theory: A nonextensional representation of subjective probability. *Psychological Review*. 101. pp. 547–67 (Reprinted in edited form in Gilovich, T., Griffin, D., and Kahneman, D. *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge: Cambridge University Press.

Winkler, R. L. (1971) Probabilistic prediction: Some experimental results. *Journal of the American Statistical Association*. 66. pp. 675–85.

Winkler, R. L. (1996) Scoring rules and the evaluation of probabilities (with discussion and reply). *Test*. 5. pp. 1–60.

Winman, A., Hansson, P., and Juslin, P. (2004) Subjective probability intervals: How to cure overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning Memory and Cognition*. 30. pp. 1167–75.

CHAPTER 25

PROBABILISTIC OPINION POOLING

FRANZ DIETRICH AND CHRISTIAN LIST

25.1 Introduction

How can several individuals' opinions on some factual matters be aggregated into unified collective opinions? This question arises in many contexts. A panel of climate experts may have to aggregate the panelists' conflicting opinions into a compromise view, in order to deliver a report to the government. A jury may have to arrive at a collective opinion on the facts of a case, despite disagreements between the jurors, so that the court can reach a verdict. In Bayesian statistics, we may wish to specify some all-things-considered prior probabilities by aggregating the subjective prior probabilities of different statisticians. In meta-statistics, we may wish to aggregate the probability estimates that different statistical studies have produced for the same events. An individual agent may wish to combine his or her own opinions with those of another, so as to resolve any peer disagreements. Finally, in a purely intra-personal case, an agent may seek to reconcile different 'selves' by aggregating their conflicting opinions on the safety of mountaineering, in order to decide whether to undertake a mountain hike and which equipment to buy.

How *should* opinions be aggregated in each of these cases? Perhaps surprisingly, this question has no obvious answer. Of course, if there is unanimity on the facts in question, we can simply take the unanimous opinions as the collective ones. But as soon as there are even mild disagreements, the aggregation problem becomes non-trivial. The aim of this chapter is to review and assess some salient proposed solutions.

Our focus will be on the aggregation of *probabilistic* opinions, which is often called the problem of *opinion pooling*. For present purposes, the opinions take the form of assignments of probabilities to some events or propositions of interest. Suppose, for instance, our climate panel consists of three experts, who assign the probabilities 0.3, 0.5, and 0.7 to the event that the global temperature will rise by more than one degree Celsius in the next 20 years. One proposal is to compute the linear average of these probabilities, so that the collective probability of the event is $\frac{1}{3}0.3 + \frac{1}{3}0.5 + \frac{1}{3}0.7 = 0.5$. Another proposal is to compute

a weighted linear average of the form $w_1 0.3 + w_2 0.5 + w_3 0.7$, where w_1, w_2, and w_3 are non-negative weights whose sum-total is 1. Each expert's weight could reflect his or her competence, so that more competent experts have greater influence on the collective opinions. If expert 1 is deemed more competent than experts 2 and 3, then w_1 may be closer to 1, while w_2 and w_3 may be closer to 0. (In the special case of equal weights, we speak of an *unweighted* average.) A third proposal is to compute a geometric, rather than linear, average of the individuals' probabilities, which could also be weighted or unweighted. Generally, a *pooling function*, defined formally below, is a function from individual to collective probability assignments. Clearly, we can define many different pooling functions; the linear, weighted linear, and geometric functions are just illustrations.

Which pooling function is appropriate depends on the context and the intended status of the collective opinions. At least three questions are relevant:

- Should the collective opinions represent a *compromise* or a *consensus*? In the first case, each individual may keep his or her own personal opinions and adopt the collective opinions only hypothetically when representing the group or acting on behalf of it. In the second case, all individuals are supposed to take on the collective opinions as their own, so that the aggregation process can be viewed as a consensus formation process.
- Should the collective opinions be justified on *epistemic* or *procedural* grounds? In the first case, the pooling function should generate collective opinions that are epistemically well-justified: they should 'reflect the relevant evidence' or 'track the truth', for example. In the second case, the collective opinions should be a fair representation of the individual opinions. The contrast between the two approaches becomes apparent when different individuals have different levels of competence, so that some individuals' opinions are more reliable than others'. The epistemic approach then suggests that the collective opinions should depend primarily on the opinions of the more competent individuals, while the procedural approach might require that all individuals be given equal weight.
- Are the individuals' opinions based only on *shared information* or also on *private information*? This, in turn, may depend on whether the group has deliberated about the subject matter before opinions are aggregated. Group deliberation may influence individual opinions as the individuals learn new information and become aware of new aspects of the issue. It may help remove interpersonal asymmetries in information and awareness.

As we will see, *linear pooling* (the weighted or unweighted linear averaging of probabilities) can be justified on procedural grounds but not on epistemic ones, despite the possibility of giving greater weight to more competent individuals. Epistemic considerations support two other pooling methods: *geometric pooling* (the weighted or unweighted geometric averaging of probabilities), and *multiplicative pooling* (where probabilities are multiplied rather than averaged). The choice between geometric and multiplicative pooling, in turn, depends on whether the individuals' opinions are based on shared information or on private information.

After setting the stage in Sections 25.2 and 25.3, we discuss linear pooling in Sections 25.4 and 25.5, geometric pooling in Sections 25.6 and 25.7, and multiplicative pooling in Sections 25.8 and 25.9. We give an axiomatic characterization of each class of pooling

functions and assess its plausibility. The characterizations are well-known in the case of linear and geometric pooling, but – to the best of our knowledge – new in the case of multiplicative pooling. In Section 25.10, we briefly mention some further approaches to opinion pooling: *supra-Bayesian pooling* (a radically Bayesian approach), the *aggregation of imprecise or qualitative probabilities*, the aggregation of binary yes/no opinions, known as *judgment aggregation*, and some generalized kinds of opinion aggregation.

There is a growing interdisciplinary literature on probabilistic opinion pooling; some references are given below (for a classic review, see Genest and Zidek 1986). While a complete review of the literature is beyond the scope of this chapter, we aim to give a flavour of the variety of possible approaches. We will discuss what we take to be the main arguments for and against the three central approaches we are focusing on: linear, geometric, and multiplicative pooling. As we will argue, these approaches promote different goals and rest on different assumptions.

25.2 The Problem of Probabilistic Opinion Pooling

We consider a group of $n \geq 2$ individuals, labelled $i = 1, \ldots, n$, who have to assign probabilities to some events.

The agenda.

The *agenda* is the set of events under consideration. We define events as sets of possible worlds. Formally, consider a fixed non-empty set Ω of *possible worlds* (sometimes also called *possible states*). We take Ω to be finite for simplicity (but almost everything we say could be generalized to the infinite case). An *event* is a subset A of Ω; it can also be interpreted as a *proposition*. The complement of any event A is denoted $A^c = \Omega \backslash A$ and can be interpreted as its *negation*. For any two events A and B, the intersection $A \cap B$ can be interpreted as their *conjunction*, and the union $A \cup B$ as their *disjunction*. The events Ω (the entire set) and $\varnothing$ (the empty set) represent the *tautology* and the *contradiction*, respectively. All other events are *contingent*. For present purposes, the agenda is simply the set of all possible events, formally the power set of Ω (the set of all subsets of Ω), denoted $2^{\Omega} = \{A : A \subseteq \Omega\}$.

The simplest non-trivial example is a set of two worlds, $\Omega = \{\omega, \omega'\}$. Here, the agenda contains only two contingent events, namely $\{\omega\}$ and $\{\omega'\}$, e.g., 'rain' and 'no rain'. Obviously, the agenda grows exponentially in the size of Ω.[1]

A concrete agenda.

As an illustration (from Dietrich and List 2013a,b), consider an expert committee that seeks to form collective opinions about climate change. Possible worlds are vectors (j, k, l) of three characteristics, which may each take the value 0 or 1:

[1] While we here take the agenda to consist of all possible events $A \subseteq \Omega$ (so that it is always closed under the Boolean operations of conjunction, disjunction, and negation), this classical but demanding assumption is dropped in Dietrich and List (2013a), where the agenda is not required to be closed under Boolean operations.

- The first characteristic specifies whether greenhouse-gas concentrations exceed some critical threshold ($j = 1$) or not ($j = 0$).
- The second characteristic specifies whether there is a causal law by which greenhouse-gas concentrations above the threshold cause Arctic summers to be ice-free ($k = 1$) or not ($k = 0$).
- The third characteristic specifies whether Arctic summers are ice-free ($l = 1$) or not ($l = 0$).

Formally, the set of possible worlds is

$$\Omega = \{(1,1,1),(1,0,1),(1,0,0),(0,1,1),(0,1,0),(0,0,1),(0,0,0)\}.$$

This is the set of all triples of 0s and 1s with the exception of $(1,1,0)$. The latter triple is excluded because it represents an inconsistent combination of characteristics. The expert committee must assign a collective probability to every event $A \subseteq \Omega$.

The opinions.

Opinions are represented by *probability functions*. A *probability function* P assigns to each event $A \subseteq \Omega$ a real number $P(A)$ in $[0,1]$ such that

- the tautology has probability one: $P(\Omega) = 1$; and
- P is additive: $P(A \cup B) = P(A) + P(B)$ whenever two events A and B are mutually inconsistent, i.e., $A \cap B = \varnothing$.

The probability of a singleton event $\{\omega\}$ is often denoted $P(\omega)$ rather than $P(\{\omega\})$. Clearly, the probability of any event A can be written as the sum $P(A) = \sum_{\omega \in A} P(\omega)$. Thus a probability function P is fully determined by the probabilities $P(\omega)$ of the different worlds ω in Ω. Let

- $\mathcal{P}$ be the set of all probability functions P, and
- $\mathcal{P}'$ be the set of all probability functions P which are *regular*, i.e., $P(\omega) > 0$ for all worlds ω.

Opinion pooling.

A combination of probability functions across the n individuals, $(P_1,\ldots,P_n)$, is called an *opinion profile*. A *pooling function* takes opinion profiles as input and produces collective probability functions as output. Formally, it is a function, F, which maps each opinion profile $(P_1,\ldots,P_n)$ within some domain of admissible profiles to a single probability function $P_{P_1,\ldots,P_n} = F(P_1,\ldots,P_n)$. The notation $P_{P_1,\ldots,P_n}$ indicates that the collective probability function depends on the individual probability functions $P_1,\ldots,P_n$.

Some pooling functions are defined on the domain of all logically possible opinion profiles, others on a more restricted domain, such as the domain of all profiles of *regular* probability functions. In the first case, F is a function from $\mathcal{P}^n$ to $\mathcal{P}$; in the second case, a function from $\mathcal{P}'^n$ to $\mathcal{P}$. (As is standard, for any set S, we write S^n to denote the set of all n-tuples consisting of elements of S.)

The linear example.

The best-known example is a *linear* pooling function, which goes back to Stone (1961) or even Laplace.[2] Here, each opinion profile $(P_1, \ldots, P_n)$ in the domain $\mathcal{P}^n$ is mapped to the collective probability function satisfying

$$P_{P_1,\ldots,P_n}(A) = w_1 P_1(A) + \cdots + w_n P_n(A) \text{ for every event } A \subseteq \Omega,$$

where $w_1, \ldots, w_n$ are fixed non-negative weights with sum-total 1. The class of linear pooling functions includes a variety of functions, ranging from linear averaging with equal weights, where $w_i = \frac{1}{n}$ for all i, to an 'expert rule' or 'dictatorship', where $w_i = 1$ for one individual and $w_j = 0$ for everyone else. In the latter case:

$$P_{P_1,\ldots,P_n}(A) = P_i(A) \text{ for every event } A \subseteq \Omega.$$

25.3 The Axiomatic Method

As should be clear, there is an enormous number of logically possible pooling functions. Many are unattractive. For example, we would not normally want the collective probability of any event to depend *negatively* on the individual probabilities of that event. (An example of a negative dependence would be a case in which the individual probabilities for some event all go up, while the collective probability goes down, with all relevant other things remaining equal.) Similarly, we would not normally want the collective probabilities to depend only on the probabilities assigned by a single 'dictatorial' individual. How can we choose a good pooling function? Here, the *axiomatic method* comes into play. Under this method, we do not choose a particular pooling function directly, say linear pooling, but instead formulate general requirements on a 'good' pooling function – our *axioms* – and then ask which pooling functions, if any, satisfy them. One example is the axiom of unanimity preservation, which requires that if all individuals hold the same opinions, these opinions become the collective ones. This is satisfied by linear pooling functions, but also by many other pooling functions. So, this axiom does not single out a unique pooling function. However, if we add another axiom, as discussed below, we can narrow down the class of possible pooling functions to the class of linear pooling functions alone.

The axiomatic method can guide and structure our search for a good pooling function. The difficult question of which pooling function to use is re-cast as the more tractable question of which axioms to impose. This allows us to assess different axioms one by one rather than having to assess a fully specified pooling function in one go.

Generally, once we have specified a set of axioms, we will be faced with one of three possible situations:

(1) *Exactly one* pooling function – or one salient class of pooling functions – satisfies all our axioms, in which case we have successfully completed our search for a pooling function.

[2] For other early contributions, see Bacharach (1972) and DeGroot (1974).

(2) *Several* pooling functions – or even several structurally different classes of pooling functions – satisfy all our axioms. This is a case of *underdetermination*, in which we may wish to impose further axioms.
(3) *No* pooling function satisfies all our axioms. This is a case of *overdetermination*, in which we may have to relax at least one axiom.

25.4 Linear Pooling: the Eventwise Independent Approach

Which axioms characterize the class of linear pooling functions? Aczél and Wagner (1980) and McConway (1981) give an elegant answer to this question, identifying two jointly necessary and sufficient axioms: *eventwise independence* and *unanimity preservation*.

The first, *eventwise independence* (or simply *independence*), requires that the collective probability of any event depend solely on the individual probabilities of *that* event.[3] This reflects the democratic idea that the collective opinion on any issue should be determined by individual opinions on that issue. The underlying picture of democracy is a *non-holistic* one, under which the collective opinion on any issue must not be influenced by individual opinions on other issues.

Independence. For each event $A \in X$, there exists a function $D_A : [0,1]^n \to [0,1]$, called the *local pooling criterion* for A, such that

$$P_{P_1,\ldots,P_n}(A) = D_A(P_1(A),\ldots,P_n(A))$$

for every opinion profile $(P_1,\ldots,P_n)$ in the domain of the pooling function.

Each local pooling criterion D_A aggregates any combination of probabilities $(x_1,\ldots,x_n)$ on a specific event into a single collective probability $D_A(x_1,\ldots,x_n)$. In the case of a linear pooling function, the local pooling criterion for any event A is simply $D_A = D$, with

$$D(x_1,\ldots,x_n) = w_1x_1 + \cdots + w_nx_n,$$

where $w_1, w_2, \ldots, w_n$ are the weights of the n individuals.

The second axiom, *unanimity preservation*, requires that if all individuals hold the same opinions, these opinions become the collective ones:

Unanimity preservation. For every opinion profile $(P_1,\ldots,P_n)$ in the domain of the pooling function, if all P_i are identical, then $P_{P_1,\ldots,P_n}$ is identical to them.

This axiom seems very compelling, especially from the procedural perspective of making collective probabilities responsive to individual probabilities. Surprisingly, however, the axiom may be problematic from an epistemic perspective (see Section 25.7), but for now we do not question it.

[3] This axiom is also known as the *weak setwise function property* or *weak label neutrality*.

Theorem 1. (Aczél and Wagner 1980; McConway 1981) Suppose $|\Omega| > 2$. The linear pooling functions are the only independent and unanimity-preserving pooling functions (with domain $\mathcal{P}^n$).[4]

This result is surprising, because eventwise independence seems, at first, to leave a great degree of freedom in the specification of the local pooling criteria D_A. In conjunction with unanimity preservation, however, independence becomes quite restrictive. First, each local pooling criterion D_A must then be a linear averaging criterion. Second, the local pooling criteria D_A must be the same for all events A. This precludes defining the collective probability for any event A as the weighted average

$$P_{P_1,\ldots,P_n}(A) = D_A(P_1(A),\ldots,P_n(A)) = w_1^A P_1(A) + \cdots + w_n^A P_n(A), \tag{25.1}$$

where an individual i may have different weights w_i^A for different events A. One might consider such event-dependent weights plausible, because an individual need not be equally good at estimating the probabilities of different events. Ideally, one might wish to give each individual a greater weight in determining the collective probability for events within his or her area of expertise than for events outside that area. Unfortunately, formula (25.1) does not guarantee a well-defined collective probability function unless each individual's weight w_i^A is the same for all events A ($\neq \Omega, \varnothing$), as in standard linear pooling. In particular, if weights vary across events, the function defined in (25.1) can violate additivity.

What can be said in defence of eventwise independence? There are at least two pragmatic arguments for it. First, eventwise independent aggregation is easy to implement, because it permits the subdivision of a complex aggregation problem into multiple simpler ones, each focusing on a single event. Our climate panel can first consider the event that greenhouse-gas concentrations exceed some critical threshold and aggregate individual probabilities for that event; then do the same for the second event; and so on. Second, eventwise independent aggregation is invulnerable to agenda manipulation. If the collective opinion about each event depends only on the individual opinions about that event, then an agenda-setter who might wish to influence the outcome of the aggregation will not be able to change the collective opinion about any event by adding further events to the agenda or removing others from it. For instance, the agenda-setter could not affect the collective probability for the event 'snow' by adding the event 'hail' to the agenda.[5] McConway (1981) proves that eventwise independence is equivalent to the requirement that collective

[4] To be precise, Aczél and Wagner (1980) and McConway (1981) use another, logically independent unanimity axiom, called *zero preservation*: if some event is assigned zero probability by each individual, then it is assigned zero probability collectively. As another alternative, one could use the following axiom, which weakens both of these conditions: if some *world* ω is assigned probability 1 by every individual (so that everyone holds the same degenerate probability function), then ω is assigned probability 1 collectively. Other axiomatic characterizations of linear pooling are given by Mongin (1995) and Chambers (2007). See also Lehrer and Wagner (1981), who use linear opinion pooling to build a theory of consensus formation in groups. Linear pooling is further analysed by Wagner (1982, 1985) and Aczél, Ng, and Wagner (1984).

[5] A change in the agenda would have to be represented mathematically by a change in the underlying set of worlds Ω. In order to add the event 'hail' to the agenda, each world ω in the original set Ω must be replaced by two worlds, ω_1 and ω_2, interpreted as ω combined with the occurrence of hail and ω combined with its non-occurrence, respectively.

opinions be invariant under changes in the specification of the agenda; see also Genest (1984b).[6]

25.5 The Limitations of Eventwise Independent Aggregation

There are a number of objections to eventwise independence and consequently to linear pooling. First, it is questionable whether eventwise independent aggregation can be justified epistemically. The collective opinions it generates may not adequately incorporate the information on which individual opinions are based. As we will see in Sections 25.6 to 25.9, some axioms that capture the idea of 'adequately incorporating information' – namely the axioms of *external Bayesianity* and *individualwise Bayesianity* – typically lead to pooling functions that violate eventwise independence.

Secondly, eventwise independence becomes implausible when this requirement is applied to 'artificial' composite events, such as conjunctions or disjunctions of intuitively unrelated events. There seems no reason, for example, why the collective probability for the disjunction 'snow *or* wind' should depend only on individual probabilities for that disjunction, rather than on individual probabilities for each disjunct. Except in trivial cases, the agenda will always contain some 'artificial' composite events, since it is closed under Boolean operations (conjunction, disjunction, and negation). (Eventwise independence may become more plausible if we relax this closure requirement on the agenda; see Dietrich and List 2013a.)

Finally, eventwise independence conflicts with the principle of *preserving probabilistic independence*. This requires that any two events that are uncorrelated according to every individual's probability function remain uncorrelated according to the collective probability function. For instance, if each climate expert took the events of high greenhouse-gas concentrations and ice-free Arctic summers to be uncorrelated, then these two events should remain uncorrelated according to the collective probabilities. Unfortunately, as shown by Wagner (1984), eventwise independent pooling functions do not preserve probabilistic independence (setting aside degenerate pooling functions, such as dictatorial ones).

In fairness, we should mention that the failure to preserve probabilistic independence can be held not just against eventwise independent pooling functions but also against a much wider class of pooling functions (Genest and Wagner 1987). This includes all linear, geometric, and multiplicative pooling functions that are non-dictatorial. Further, the preservation of probabilistic independence is itself a normatively questionable requirement. Why, for example, should probabilistic independence judgments be preserved even when they are purely accidental, i.e., not driven by any insight into the causal connections

[6] McConway captures this requirement by the so-called *marginalization property*, which requires aggregation to commute with the operation of reducing the relevant algebra (agenda) to a sub-algebra (sub-agenda); this reduction corresponds to the removal of events from the agenda.

between events? It is more plausible to require that only *structurally relevant* probabilistic independencies be preserved, i.e., those that are due to the structure of causal connections rather than being merely accidental. On the preservation of causally motivated probabilistic independencies, see Bradley, Dietrich, and List (2014).

25.6 Geometric Pooling: the Externally Bayesian Approach

We now turn to a class of pooling functions based on geometric, rather than linear, averaging. While the linear average of n numbers, such as $x_1, x_2, \ldots, x_n$, is $\frac{x_1+x_2+\cdots+x_n}{n}$, the geometric average is $\sqrt[n]{x_1 x_2 \cdots x_n} = x_1^{\frac{1}{n}} x_2^{\frac{1}{n}} \cdots x_n^{\frac{1}{n}}$. Just as a linear average can be generalized to take the weighted form $w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$, so a geometric average can be generalized to take the weighted form $x_1^{w_1} x_2^{w_2} \cdots x_n^{w_n}$, where $w_1, \ldots, w_n$ are non-negative weights with sum-total 1.

A geometric pooling function determines the collective probabilities in two steps. In the first step, it takes the collective probability of each possible world (rather than event) to be a geometric average of the individuals' probabilities of that world. In the second step, it renormalizes these collective probabilities in such a way that their sum total becomes 1.

Formally, a pooling function is called *geometric* (or also *logarithmic*) if it maps each opinion profile $(P_1, \ldots, P_n)$ in the domain $\mathcal{P}'^n$ to the collective probability function satisfying

$$P_{P_1,\ldots,P_n}(\omega) = c[P_1(\omega)]^{w_1} \cdots [P_n(\omega)]^{w_n} \text{ for every world } \omega \text{ in } \Omega,$$

where $w_1, \ldots, w_n$ are fixed non-negative weights with sum total 1 and c is a normalization factor, given by

$$c = \frac{1}{\sum_{\omega' \in \Omega} [P_1(\omega')]^{w_1} \cdots [P_n(\omega')]^{w_n}}.$$

The sole point of the normalization factor c is to ensure that the sum total of the collective probabilities across all worlds in Ω becomes 1. Two technical points are worth noting. First, geometric pooling functions are defined by specifying the collective probabilities of worlds, rather than events, but this is of course sufficient to determine the collective probabilities of all events. Second, to ensure well-definedness, the domain of a geometric pooling function must be $\mathcal{P}'^n$ rather than $\mathcal{P}^n$, admitting only *regular* individual probability functions as input.[7]

As in the case of linear pooling, geometric pooling functions can be weighted or unweighted, ranging from geometric averaging with equal weights, where $w_i = \frac{1}{n}$ for all i,

[7] Without this restriction, it could happen that, for every world, *some* individual assigns a probability of zero to it, so that the geometric average of individual probabilities is zero for all worlds, a violation of well-definedness. A similar remark applies to the definition of multiplicative pooling in the next section.

to an 'expert rule' or 'dictatorship', where $w_i = 1$ for one individual and $w_j = 0$ for everyone else, so that $P_{P_1,\ldots,P_n} = P_i$.

How can geometric pooling be justified? First, it clearly preserves unanimity. Second, unlike linear pooling, it is not eventwise independent (except in the limiting case of an expert rule or dictatorship). Intuitively, this is because the renormalization of probabilities introduces a holistic element.

However, geometric pooling satisfies another, epistemically motivated axiom, called *external Bayesianity* (proposed by Madansky 1964). This concerns the effects that informational updating has on individual and collective probability functions. Informally, the axiom requires that, if probabilities are to be updated based on some information, it should make no difference whether they are updated *before* aggregation or *after* aggregation. We should arrive at the same collective probability function irrespective of whether the individuals first update their probability functions and then aggregate them, or whether they first aggregate their probability functions and then update the resulting collective probability function, where the update is based on the same information.

To formalize this, we represent information by a *likelihood function*. This is a function L which assigns, to each world ω in Ω, a positive number $L(\omega)$, interpreted as the degree to which the information supports ω, or more precisely the likelihood that the information is true in world ω. In our climate-panel example, the information that a revolutionary carbon-capture-and-storage technology is in use may be expressed by a likelihood function L that takes lower values at worlds with high greenhouse-gas concentrations than at worlds with low greenhouse-gas concentrations. This is because the information is more likely to be true at worlds with low greenhouse-gas concentrations than at worlds with high ones. (The revolutionary carbon-capture-and-storage technology would remove greenhouse-gases from the atmosphere.)

What does it mean to update a probability function based on the likelihood function L? Suppose an agent initially holds the probability function P and now learns the information represented by L. Then the agent should adopt the new probability function P^L satisfying

$$P^L(\omega) := \frac{P(\omega)L(\omega)}{\sum_{\omega' \in \Omega} P(\omega')L(\omega')} \text{ for every world } \omega \text{ in } \Omega. \tag{25.2}$$

This definition can be motivated in Bayesian terms. For a simple illustration, consider a limiting case of a likelihood function L, where $L(\omega) = 1$ for all worlds ω within some event A and $L(\omega) = 0$ for all worlds ω outside A. Here L simply expresses the information that event A has occurred. (This is a limiting case of a likelihood function because L is not positive for all worlds ω, as required by our definition, but only non-negative.) Formula (25.2) then reduces to the familiar requirement that the agent's posterior probability function after learning that event A has occurred be equal to his or her prior probability function conditional on A. In the Appendix, we discuss the notion of a likelihood function and the Bayesian motivation for formula (25.2) in more detail.

The axiom of external Bayesianity can now be stated as follows:

External Bayesianity. For every opinion profile $(P_1, \ldots, P_n)$ in the domain of the pooling function and every likelihood function L (where the updated profile $(P_1^L, \ldots, P_n^L)$ remains in the domain), pooling and updating are commutative, i.e., $P_{P_1^L,\ldots,P_n^L} = P^L_{P_1,\ldots,P_n}$.

Theorem 2. (e.g. Genest 1984a) The geometric pooling functions are externally Bayesian and unanimity-preserving.

Let us briefly explain why a geometric pooling function (say with weights $w_1, \ldots, w_n$) is externally Bayesian. Without loss of generality, we can view any probability function as a function from the set Ω of worlds into $[0,1]$, rather than as a function from the set 2^Ω of events into $[0,1]$. Consider any opinion profile $(P_1, \ldots, P_n)$ (in the domain $\mathcal{P}'^n$) and any likelihood function L. To show that $P_{P_1^L,\ldots,P_n^L} = P^L_{P_1,\ldots,P_n}$, we observe that each side of this equation is proportional to the function $[P_1]^{w_1} \cdots [P_n]^{w_n} L$. (Since we are dealing with probability functions, proportionality then implies identity.) First, note that $P^L_{P_1,\ldots,P_n}$ is proportional to this function by definition. Second, note that $P_{P_1^L,\ldots,P_n^L}$ is proportional to the product of functions $[P_1^L]^{w_1} \cdots [P_n^L]^{w_n}$, also by definition. But, since each function P_i^L is proportional to the product $P_i L$, the product $[P_1^L]^{w_1} \cdots [P_n^L]^{w_n}$ is, in turn, proportional to the function

$$[P_1 L]^{w_1} \cdots [P_n L]^{w_n} = [P_1]^{w_1} \cdots [P_n]^{w_n} L^{w_1 + \cdots + w_n} = [P_1]^{w_1} \cdots [P_n]^{w_n} L,$$

as required.

Why is external Bayesianity a plausible requirement? If it is violated, the time at which an informational update occurs can influence the collective opinions. It will then matter whether the informational update takes place before or after individual opinions are aggregated. This would open the door to manipulation of the collective opinions by someone who strategically discloses a relevant piece of information at the right time. Of course, someone acting in this way need not have bad intentions; he or she might simply wish to 'improve' the collective opinions. Nonetheless, the need to decide whether $P_{P_1^L,\ldots,P_n^L}$ or $P^L_{P_1,\ldots,P_n}$ is a 'better' collective probability function raises all sorts of complications, which we can avoid if external Bayesianity is satisfied. See Raiffa (1968, pp. 221–6) for some examples of strategic information retention when external Bayesianity is violated.

Geometric pooling functions are not the only externally Bayesian and unanimity-preserving pooling functions. The two axioms are also compatible with a generalized form of geometric pooling, in which the weights $w_1, \ldots, w_n$ may depend on the opinion profile $(P_1, \ldots, P_n)$ in a systematic way.[8] Genest, McConway, and Schervish (1986) characterize all pooling functions satisfying the conditions of Theorem 2, or just external Bayesianity. Once some additional axioms are imposed, over and above those in Theorem 2, geometric pooling becomes unique (Genest 1984a; Genest, McConway, and Schervish 1986). However, the additional axioms are technical and arguably not independently compelling. So, we still lack a fully compelling axiomatic characterization of geometric pooling. For a further discussion and comparison of linear and geometric pooling, see Genest and Zidek (1986).

[8] Let us write $w_i^{P_1,\ldots,P_n}$ for individual i's weight when the profile is $(P_1, \ldots, P_n)$. In the profile-dependent specification of weights, all one needs to ensure is that, for all i, $w_i^{P_1,\ldots,P_n} = w_i^{P'_1,\ldots,P'_n}$ whenever the profile $(P'_1, \ldots, P'_n)$ is 'accessible via update' from the profile $(P_1, \ldots, P_n)$ in the sense that there is a likelihood function L such that $P_i^L = P'_i$ for every i. Accessibility via updates defines an equivalence relation between profiles in $\mathcal{P}'^n$. Since there are many equivalence classes (provided $|\Omega| > 1$), there are many generalized geometric pooling functions.

25.7 From Symmetric to Asymmetric Information

Although we have justified geometric pooling in epistemic terms – by invoking the axiom of external Bayesianity – there are conditions under which geometric pooling is not epistemically justified. These conditions motivate another approach to opinion pooling, called *multiplicative pooling* (Dietrich 2010), which we introduce in the next section. To identify those conditions, we must consider not just the probability functions P_1, P_2, …, P_n that are to be pooled, but their *informational bases*: the information that the individuals have used to arrive at them.

Let us contrast two diametrically opposed cases, setting aside any intermediate cases for simplicity. (We comment briefly on intermediate cases in Section 25.10.)

Case 1: informational symmetry. The individuals' probability functions $P_1, \ldots, P_n$ are based on exactly the same information. Any differences in these probability functions stem at most from different ways of interpreting that shared information.

Case 2: informational asymmetry. The individuals' probability functions P_1, …, P_n are based on different information, and there is no overlap between different individuals' information, apart from some fixed background information held by everyone. Each individual i's probability function P_i is derived from some prior probability function by conditionalizing on i's private information. That is, $P_i = p_i^{L_i}$, where p_i is i's prior probability function and L_i is the likelihood function representing i's private information. For simplicity, we assume a shared prior probability function $p_i = p$ for every individual i, which reflects the individuals' shared background information.

Case 1 might occur if there is group deliberation and exchange of information prior to the pooling of opinions. Case 2 might occur in the absence of such group deliberation or exchange of information. We will now show that the axioms by which we have justified geometric pooling – unanimity preservation and external Bayesianity – are plausible in Case 1, but not in Case 2.

Consider unanimity preservation. In Case 1, this axiom is compelling. If all individuals arrive at the same probability function $P_1 = \ldots = P_n$ based on shared information, there is no reason why this probability function should not also become the collective one. After all, in the present case, the individuals not only *have* the same information, as assumed in Case 1, but also *interpret* it in the same way; otherwise, we would not have $P_1 = \ldots = P_n$.

In Case 2, by contrast, unanimity preservation is not compelling. If all individuals arrive at the same probability function P_i based on different private information, the collective probability function ought to incorporate that dispersed information. Thus it should incorporate the individuals' likelihood functions $L_1, \ldots, L_n$, and this may, in turn, require a collective probability function distinct from $P_1 = \ldots = P_n$.[9] Suppose, for example, that all the experts on our climate panel assign the same high probability of 0.9 to the

[9] One may want to obtain the collective probability function $P_{P_1,\ldots,P_n}$ by updating some prior probability function p in light of all n likelihood functions. Then $P_{P_1,\ldots,P_n}$ equals $(\ldots((p^{L_1})^{L_2})\ldots)^{L_n}$,

event that greenhouse-gas concentrations exceed the critical threshold. Plausibly, if each expert has some private information that supports assigning a high probability to some event, compared to a much lower prior, then the totality of private information supports the assignment of an even higher probability to it. Thus the collective probability should not be the same as the individual ones, but amplified, above 0.9. Similarly, if all experts, prompted by their own independent evidence, assign the same low probability of 0.1 to some event, then the collective probability should be even lower. Here, the group knows more than each individual member.

Next consider external Bayesianity. In Case 1, where all individuals have the same information, this requirement is well motivated, as should be clear from our discussion in the last section. By contrast, in Case 2, where different individuals have different and non-overlapping private information, external Bayesianity loses its force. Recall that we justified the requirement that $P_{P_1^L,\ldots,P_n^L} = P^L_{P_1,\ldots,P_n}$ by interpreting L as representing information that is received by *all* individuals. In Case 2, however, individuals have only private information (apart from some shared but fixed background information, which cannot include the non-fixed information represented by L).[10] Here, updating all probability functions P_i would mean updating them *on the basis of different private information*. So, the updated profile $(P_1^L, \ldots, P_n^L)$ would have to be interpreted as expressing the individuals' opinions after incorporating *different* items of private information that happen to be represented by the same likelihood function L for each individual. This interpretation makes it implausible to require that $P_{P_1^L,\ldots,P_n^L}$ and $P^L_{P_1,\ldots,P_n}$ be the same. From the group's perspective, there is not just one item of information to take into account, but n separate such items. While each item of information by itself corresponds to the likelihood function L, the group's information as a whole corresponds to the product of n such functions, namely L^n. In the next section, we introduce an axiom that replaces external Bayesianity in Case 2.

25.8 Multiplicative Pooling: the Individualwise Bayesian Approach

We now consider a class of pooling functions that are appropriate in Case 2, where the probability functions P_1, P_2, …, P_n are based on different private information and there is at most some fixed background information held by all individuals. This is the class of multiplicative pooling functions (proposed by Dietrich 2010), which are based on multiplying, rather than averaging, probabilities.

which in turn equals $p^{L_1 L_2 \cdots L_n}$, the probability function obtained by updating p in light of the likelihood function defined as the product $L_1 L_2 \cdots L_n$. This is, in effect, what multiplicative pooling does, as should become clear in the next section.

[10] The information represented by L is non-fixed, since it is present in one opinion profile, $(P_1^L, \ldots, P_n^L)$, and absent in another, $(P_1, \ldots, P_n)$.

A multiplicative pooling function, like a geometric one, determines the collective probabilities in two steps. In the first step, it takes the collective probability of each possible world to be the product of the individuals' probabilities of that world, calibrated by multiplication with some exogenously fixed probability (whose significance we discuss in Section 25.9). This differs from the first step of geometric pooling, where the geometric average of the individuals' probabilities is taken. In the second step, multiplicative pooling renormalizes the collective probabilities such that their sum total becomes 1; this matches the second step of geometric pooling.

Formally, a pooling function is called *multiplicative* if it maps each opinion profile $(P_1, \ldots, P_n)$ in the domain $\mathcal{P}'^n$ to the collective probability function satisfying

$$P_{P_1,\ldots,P_n}(\omega) = cP_0(\omega)P_1(\omega)\cdots P_n(\omega) \text{ for every world } \omega \text{ in } \Omega,$$

where P_0 is some fixed probability function, called the *calibrating function*, and c is a normalization factor, given by

$$c = \frac{1}{\sum_{\omega' \in \Omega} P_0(\omega')P_1(\omega')\cdots P_n(\omega')}.$$

As before, the point of the normalization factor c is to ensure that the sum total of the collective probabilities across all worlds in Ω is 1. To see that multiplicative pooling can be justified in Case 2, we now introduce a new axiom that is plausible in that case – *individualwise Bayesianity* – and show that it is necessary and sufficient for multiplicative pooling. (The present characterization of multiplicative pooling is distinct from the one given in Dietrich 2010.)

The axiom says that it should make no difference whether some information is received by a single individual before opinions are pooled or by the group as a whole afterwards. More specifically, we should arrive at the same collective probability function irrespective of whether a single individual first updates his or her own probability function based on some private information and the probability functions are then aggregated, or whether the probability functions are first aggregated and then updated – now at the collective level – given the same information.

Individualwise Bayesianity. For every opinion profile $(P_1, \ldots, P_n)$ in the domain of the pooling function, every individual i, and every likelihood function L (where the profile $(P_1, \ldots, P_i^L, \ldots, P_n)$ remains in the domain), we have $P_{P_1,\ldots,P_i^L,\ldots,P_n} = P^L_{P_1,\ldots,P_n}$.

Just as external Bayesianity was plausible in Case 1, where all individuals' probability functions are based on the same information, so individualwise Bayesianity is plausible in Case 2, where different individuals' probability functions are based on different private information. The argument for individualwise Bayesianity mirrors that for external Bayesianity: any violation of the axiom implies that it makes a difference whether someone acquires private information *before* opinions are pooled or acquires the information and shares it with the group *afterwards*. This would again generate opportunities for manipulation by third parties able to control the acquisition of information.

Theorem 3. The multiplicative pooling functions are the only individualwise Bayesian pooling functions (with domain $\mathcal{P}'^n$).

This (new) result has an intuitive proof, which we now give.

Proof: Let us again view any probability function as a function from the set Ω of worlds into $[0,1]$, rather than as a function from the set 2^{Ω} of events into $[0,1]$. As noted earlier, this is no loss of generality. We first prove that multiplicative pooling functions satisfy individualwise Bayesianity. Consider a multiplicative pooling function, for some exogenously fixed probability function P_0, which serves as the calibrating function. Note that, for any opinion profile $(P_1, \ldots, P_n)$,

- the function $P_{P_1,\ldots,P_i^L,\ldots,P_n}$ is by definition proportional to the product $P_0 P_1 \cdots (P_i L) \cdots P_n$, and
- the function $P^L_{P_1,\ldots,P_n}$ is by definition proportional to the product $(P_0 P_1 \cdots P_n)L$.

These two products are obviously the same, so individualwise Bayesianity is satisfied. Conversely, we prove that no pooling functions other than multiplicative ones satisfy the axiom. Consider any pooling function with domain $\mathcal{P}'^n$ that satisfies individualwise Bayesianity. Let P^* be the uniform probability function, which assigns the same probability to every world in Ω. We show that our pooling function is multiplicative with calibrating function $P_0 = P_{P^*,\ldots,P^*}$. Consider any opinion profile $(P_1, \ldots, P_n)$ (in $\mathcal{P}'^n$). The argument proceeds in n steps. It could be re-stated more formally as an inductive proof.

- *Step 1*: First, consider the likelihood function $L := P_1$. The function $P_{P_1,P^*,\ldots,P^*}$ is equal to $P_{(P^*)^L,P^*,\ldots,P^*}$. By individualwise Bayesianity, this is equal to $P^L_{P^*,\ldots,P^*}$, which is in turn proportional to $P_{P^*,\ldots,P^*}L = P_0 P_1$, by the definitions of P_0 and L.
- *Step 2*: Now, consider the likelihood function $L := P_2$. The function $P_{P_1,P_2,P^*,\ldots,P^*}$ is equal to $P_{P_1,(P^*)^L,P^*,\ldots,P^*}$. By individualwise Bayesianity, this is equal to $P^L_{P_1,P^*,\ldots,P^*}$, which is in turn proportional to $P_{P_1,P^*,\ldots,P^*}L$, i.e., to $P_0 P_1 P_2$, by Step 1 and the definition of L.

  ...
- *Step n*: Finally, consider the likelihood function $L := P_n$. The function $P_{P_1,\ldots,P_n}$ is equal to $P_{P_1,\ldots,P_{n-1},(P^*)^L}$. By individualwise Bayesianity, this is equal to $P^L_{P_1,\ldots,P_{n-1},P^*}$, which is in turn proportional to $P_{P_1,\ldots,P_{n-1},P^*}L$, i.e., to $P_0 P_1 \cdots P_n$, by Step $n-1$ and the definition of L. ■

25.9 How to Calibrate a Multiplicative Pooling Function

Recall that the definition of a multiplicative pooling function involves a calibrating probability function P_0. The collective probability of each possible world is not merely the renormalized product of the individuals' probabilities of that world, but it is multiplied further by the probability that P_0 assigns to the world. How should we choose that calibrating probability function?

It is simplest to take P_0 to be the uniform probability function, which assigns the same probability to every world in Ω. In this case, we obtain the *simple multiplicative pooling function*, which maps each opinion profile $(P_1, \ldots, P_n)$ in $\mathcal{P}'^n$ to the collective probability function satisfying

$$P_{P_1,\ldots,P_n}(\omega) = cP_1(\omega)\cdots P_n(\omega) \text{ for every world } \omega \text{ in } \Omega,$$

for a suitable normalization factor c.

The simple multiplicative pooling function is the only multiplicative pooling function that satisfies an additional axiom, which we call *indifference preservation*. It is a weak version of the unanimity-preservation axiom, which applies only in the special case in which every individual's probability function is the uniform one.

Indifference preservation. If every probability function in the opinion profile $(P_1, \ldots, P_n)$ is the uniform probability function, then the collective probability function $P_{P_1,\ldots,P_n}$ is also the uniform one (assuming the profile is in the domain of the pooling function).

Corollary of Theorem 3. The simple multiplicative pooling function is the only individual-wise Bayesian and indifference-preserving pooling function (with domain $\mathcal{P}'^n$).

When is indifference preservation plausible? We suggest that it is plausible if the individuals have no shared background information at all; all their information is private. Recall that we can view each individual i's probability function P_i as being derived from a shared prior probability function p by conditionalizing on i's private information L_i. If the individuals have no shared background information, it is plausible to take p to be the uniform prior, following the *principle of insufficient reason* (though, of course, that principle raises some well-known philosophical issues, which we cannot discuss here). Any deviations from the uniform probability function on the part of some individual – i.e., in some function P_i – must then plausibly be due to some private information. But now consider the opinion profile $(P_1, \ldots, P_n)$ in which *every* P_i is the uniform probability function. For the individuals to arrive at this opinion profile, there must be a complete lack of private information, in addition to the lack of collectively shared background information. (If some individuals had relevant private information, some P_i would arguably have to be distinct from the uniform probability function.[11]) In such a situation of no information – private or shared – it seems plausible to require the collective probability function to be uniform. So, indifference preservation is plausible here.

By contrast, if the individuals have some shared background information, indifference preservation is questionable. The individuals' prior probability functions will not normally be uniform in this case, so any uniformity in an individual's posterior probability function P_i points towards the presence of some private information which has led the individual to update his or her probabilities from the non-uniform prior ones to uniform posterior ones.

[11] Alternatively, it is possible for an individual to have multiple pieces of private information that perfectly cancel each other out, so that, on balance, his or her probability function remains uniform. Strictly speaking, to justify indifference preservation in such a case, we must assume that different individuals' private information is mutually independent. We briefly discuss the issue of correlated private information in Section 25.10.

The collective probability function should therefore incorporate both the group's shared background information and the individuals' private information. There is no reason to expect that incorporating all this information will generally lead to the uniform probability function. Consequently, indifference preservation is not plausible here.

How should we choose the calibrating probability function P_0 when we cannot assume indifference preservation? Our answer to this question follows Dietrich (2010). Again, consider Case 2, where different individuals have different private information and there is at most some fixed background information that is collectively shared. Let p be every individual's prior probability function, assuming a shared prior (which may reflect the shared background information).

If none of the individuals holds any additional private information, then each individual i's probability function is simply $P_i = p$, and it is reasonable to require the group to have the same probability function p, because no further information is available to the group. Formally, $P_{p,\ldots,p} = p$.[12] By the definition of multiplicative pooling, the collective probability function $P_{p,\ldots,p}$ is proportional to the product $P_0 p^n$ (where probability functions are viewed as functions defined on the set of worlds Ω). So, p, which is equal to $P_{p,\ldots,p}$, must be proportional to $P_0 p^n$, which implies that P_0 must be proportional to $1/p^{n-1}$. Formally,

$$P_0(\omega) = \frac{c}{[p(\omega)]^{n-1}} \text{ for every world } \omega \text{ in } \Omega,$$

where c is a normalization factor to ensure that P_0 is a probability function.

This shows that the choice of P_0 is not free, but constrained by the individuals' prior probabilities. In particular, the probability assignments made by P_0 must depend strongly negatively on the individuals' prior probabilities. This idea can be generalized to the case in which different individuals have different priors, as shown in the Appendix.

25.10 Concluding Remarks

We have discussed three classes of opinion pooling functions – linear, geometric, and multiplicative – and have shown that they satisfy different axioms and are justifiable under different conditions. Linear pooling may be justified on procedural grounds, but not on epistemic grounds. Geometric and multiplicative pooling may be justified on epistemic grounds, but which of the two is appropriate depends not just on the opinion profiles to be aggregated but also on the information on which they are based. Geometric pooling can be justified if all individuals' opinions are based on the same information (Case 1), while multiplicative pooling can be justified if every individual's opinions are based solely on private information, except for some shared background information held by everyone (Case 2).

There are, of course, many intermediate cases between Case 1 and Case 2, in which the opinion pooling problem becomes more complicated. First, there are cases in which an

[12] To ensure that the opinion profile $(p,\ldots,p)$ is in the domain of the pooling function, we assume that p belongs to $\mathcal{P}'$, i.e., is a regular probability function.

opinion profile is based on some information that is neither shared by everyone, nor held by a single individual alone, but shared by a proper subset of the individuals. In such cases, neither geometric nor multiplicative pooling is justified but a more complicated pooling function – involving a recursive construction – is needed (see Dietrich 2010).

Secondly, there are cases in which there are *correlations* between different individuals' private information – a possibility implicitly assumed away in our discussion so far. If different individuals' private information is correlated, the axiom of individualwise Bayesianity loses its force. To see this, note that the combined evidential strength of two pieces of *correlated* private information, represented by the likelihood functions L_1 and L_2, is not their product $L_1 L_2$. So, it is not plausible to demand that $P_{P_1,\ldots,P_i^{L_1},\ldots,P_j^{L_2},\ldots,P_n} = P^{L_1 L_2}_{P_1,\ldots,P_n}$, as individualwise Bayesianity (applied twice) would require. (On the subject of dependencies between different individuals' opinions, see Dietrich and List 2004 and Dietrich and Spiekermann 2013.)

In sum, it should be clear that there is no one-size-fits-all approach to probabilistic opinion pooling. We wish to conclude by mentioning some other approaches that we have not discussed. One such approach is *supra-Bayesian opinion pooling* (introduced by Morris 1974), a radically Bayesian approach. Here, the collective probability of each possible world is defined as the *posterior* probability of that world (held by a hypothetical Bayesian observer), conditional on learning what the opinion profile is. Opinion pooling then becomes a complex form of Bayesian updating. This presupposes a very rich probability model, which specifies not just the prior probability of each possible world, but also the probability of obtaining each possible opinion profile conditional on each possible world. In practice, it is unclear where such a rich model could come from, and how a group could agree on it. Nevertheless, from a radically Bayesian perspective, supra-Bayesian pooling is a very natural approach – or even the rationally required one.

There are also a number of approaches that not only lead to different opinion pooling functions but redefine the aggregation problem itself. Here, the opinions to be aggregated are no longer given by probability functions, but by other formal objects. Two examples are the aggregation of *imprecise probabilities* (e.g., Moral and Sagrado 1998) and the aggregation of *ordinal probabilities*, which are expressed by probability orders (using the binary relation 'at least as probable as') rather than probability functions (e.g., Weymark 1997). Similarly, one could in principle use the tools of formal aggregation theory to study the aggregation of *ranking functions* (as discussed, e.g., by Spohn 2012).

In recent years, there has been much work on the aggregation of binary opinions, where a group seeks to assign the values 'true'/'false' or 'yes'/'no' to a set of propositions, based on the individuals' assignments – a problem now known as *judgment aggregation* (e.g., List and Pettit 2002; Dietrich 2007; Dietrich and List 2007; Nehring and Puppe 2010; Dokow and Holzman 2010; for a recent review, see List 2012). Truth-value assignments, especially in classical propositional logic, can be viewed as degenerate probability assignments (restricted to the values 0 and 1). Interestingly, the analogues of the axioms characterizing linear averaging in probabilistic opinion pooling typically lead to dictatorial aggregation in judgment-aggregation problems (for discussion, see Dietrich and List 2010).

Pauly and van Hees (2006) consider judgment-aggregation problems in many-valued (as distinct from two-valued) logics and show that some of the dictatorship results familiar from the two-valued case continue to hold in the many-valued case (for further results, see

Duddy and Piggins 2013). Relatedly, Bradley and Wagner (2012) discuss the aggregation of probability functions that take values within a finite grid, such as the grid $\{k/10 : k = 0, 1, \ldots, 10\}$. They show that this aggregation problem is also susceptible to dictatorship results akin to those in judgment aggregation. Under certain conditions, the only eventwise independent and unanimity-preserving aggregation functions are the dictatorial ones.

The list of examples could be continued. For a unified framework that subsumes several aggregation problems under the umbrella of *attitude aggregation*, see Dietrich and List (2010). In an attitude-aggregation problem, each individual i holds an *attitude function* A_i, which assigns to each proposition on some agenda a value in some set V of admissible values, which could take a variety of forms. We must further specify some criteria determining when an attitude function counts as *consistent* or *formally rational*, and when not. The task, then, is to map each profile $(A_1, \ldots, A_n)$ of individual attitude functions in some domain to a collective attitude function. It should be evident that probabilistic opinion pooling, two-valued and many-valued judgment aggregation, and finite-grid probability aggregation can all be viewed as special cases of such attitude-aggregation problems, for different specifications of (i) the value set V and (ii) the consistency or rationality criteria. (An extension of this line of research, using an algebraic framework, can be found in Herzberg 2015.)

Finally, much of the literature on opinion pooling is inspired, at least in part, by Arrow's pioneering work in social choice theory (Arrow 1951/1963). Social choice theory, in the most general terms, addresses the aggregation of potentially conflicting individual inputs into collective outputs (for a survey, see List 2013). Much of the work in this area, following Arrow, focuses on the aggregation of preferences, welfare, or interests. The theory of opinion pooling can be seen as an epistemically oriented counterpart of Arrovian social choice theory.

Acknowledgments

We are very grateful to John Cusbert and Alan Hájek for helpful written comments on this article. Franz Dietrich gratefully acknowledges financial support from the French Agence Nationale de la Recherche (ANR-12-INEG-0006-01). Christian List gratefully acknowledges financial support from the Leverhulme Trust (via a Leverhulme Major Research Fellowship).

Appendix

Likelihood Functions and their Bayesian Interpretation

According to our definition, a *likelihood function* L assigns, to each world ω in Ω, a positive number $L(\omega)$, interpreted as the likelihood that the information is true in world ω. This notion of a likelihood function is slightly non-standard, because in statistics a likelihood function is usually associated with some information (data) that is explicitly representable in the relevant model. In our climate-panel example, by contrast, the information that a

revolutionary carbon-capture-and-storage technology is in use cannot be represented by any event $A \subseteq \Omega$. To relate our notion of a likelihood function to the more standard one, we need to make the following construction.

Let us 'split' each world $\omega = (j,k,l)$ in Ω into two more refined worlds: $\omega^+ = (j,k,l,1)$ and $\omega^- = (j,k,l,0)$, in which the fourth characteristic specifies whether or not a revolutionary carbon-capture-and-storage technology is in use. The refined set of worlds, Ω', now consists of all such 'four-dimensional' worlds, formally, $\Omega' = \Omega \times \{0,1\}$. The information that a revolutionary carbon-capture-and-storage technology is in use can then be represented as an event relative to the refined set of worlds Ω', namely the event consisting of all refined worlds whose fourth characteristic is 1; call this event E.

Under this construction, the non-standard likelihood function L on Ω corresponding to this information becomes a standard likelihood function relative to our refined set Ω'. Formally, for any unrefined world $\omega \in \Omega$,

$$L(\omega) = \Pr(E|\omega) = \frac{\Pr(\omega^+)}{\Pr(\omega)},$$

where Pr is a probability function for the refined set of worlds Ω', and any unrefined world ω in Ω is re-interpreted as the event $\{\omega^+,\omega^-\} \subseteq \Omega'$.

One can think of Pr as a refinement of a probability function for the original set Ω. Of course, different individuals i may hold different probability functions P_i on Ω, and so they may hold different refined probability functions $\Pr_i$ on Ω'. Nonetheless, the likelihood function $L(\omega) = \Pr_i(E|\omega)$ is supposed to be the same for all individuals i, as we focus on *objective* (or at least *intersubjective*) information, which has an uncontroversial interpretation in terms of its evidential support for worlds in Ω.

For present purposes, the individuals may disagree about prior probabilities, but not about the evidential value of the incoming information. A paradigmatic example of objective information is given by the case in which worlds in Ω correspond to rival statistical hypotheses (e.g., possible probabilities of 'heads' for a given coin) and the information consists of statistical data (e.g., a sequence of coin tosses).

Finally, we show that our rule for updating a probability function P based on the likelihood function L – formula (25.2) in the main text – is an instance of ordinary Bayesian updating, applied to the refined model. Note that $P^L(\omega)$, the probability assigned to ω after learning the information represented by L, can be interpreted as $\Pr(\omega|E)$, where E is the event in Ω' that corresponds to the information. By Bayes's theorem,

$$\Pr(\omega|E) = \frac{\Pr(\omega)\Pr(E|\omega)}{\sum_{\omega' \in \Omega} \Pr(\omega')\Pr(E|\omega')},$$

which reduces to

$$P^L(\omega) = \frac{P(\omega)L(\omega)}{\sum_{\omega' \in \Omega} P(\omega')L(\omega')},$$

as in formula (25.2).

How to Calibrate a Multiplicative Pooling Function When there is no Shared Prior

For each individual i, let p_i denote i's prior probability function, and let p denote the prior probability function that the group as a whole will use, without asking – for the

moment – where p comes from. Plausibly, in the absence of any private information, when each individual i's probability function is simply $P_i = p_i$, the group should stick to its own prior probability function p. Formally, $P_{p_1,\ldots,p_n} = p$. [13] By the definition of multiplicative pooling, $P_{p_1,\ldots,p_n}$ is proportional to the product $P_0 p_1 \cdots p_n$ (where probability functions are again viewed as functions defined on the set of worlds Ω). So, p, which is equal to $P_{p_1,\ldots,p_n}$, must be proportional to $P_0 p_1 \cdots p_n$, which implies that P_0 must be proportional to $p/(p_1 \cdots p_n)$. Formally,

$$P_0(\omega) = \frac{cp(\omega)}{p_1(\omega) \cdots p_n(\omega)} \text{ for every world } \omega \text{ in } \Omega, \tag{25.3}$$

where c is an appropriate normalization factor.

This expression still leaves open how to specify p, the group's prior probability function. Plausibly, it should reflect the individual prior probability functions $p_1, \ldots, p_n$. Since the individuals' prior probabilities are not based on any informational asymmetry – they are, by assumption, based on the same background information – their aggregation is an instance of Case 1. Hence, geometric pooling is a reasonable candidate for determining p on the basis of $p_1, \ldots, p_n$. If we further wish to treat the individuals equally – perhaps because we equally trust their abilities to interpret the shared background information correctly – we might use *unweighted* geometric pooling, i.e., take p to be proportional to $p_1^{1/n} \cdots p_n^{1/n}$. As a result, expression (25.3) reduces to the following general formula:

$$P_0(\omega) = \frac{c[p_1(\omega)]^{1/n} \cdots [p_n(\omega)]^{1/n}}{p_1(\omega) \cdots p_n(\omega)} = \frac{c}{[p_1(\omega) \cdots p_n(\omega)]^{1-1/n}} \text{ for every world } \omega \text{ in } \Omega,$$

where c is an appropriate normalization factor.

We have now arrived at a unique solution to our opinion pooling problem, having specified a multiplicative pooling function without any free parameters. However, the present solution is quite informationally demanding. In particular, it requires knowledge of the individuals' prior probabilities. For more details, see Dietrich (2010).

References

Aczél, J. and Wagner, C. (1980) A characterization of weighted arithmetic means. *SIAM Journal on Algebraic and Discrete Methods*. 1. 3. pp. 259–60.

Aczél, J., Ng, C. T., and Wagner, C. (1984) Aggregation theorems for allocation problems. *SIAM Journal on Algebraic and Discrete Methods*. 5. 1. pp. 1–8.

Arrow, K. J. (1951/1963) *Social Choice and Individual Values*. New York, NY: Wiley.

Bacharach, M. (1972) Scientific disagreement. Unpublished manuscript.

Bradley, R., Dietrich, F., and List, C. (2014) Aggregating causal judgments. *Philosophy of Science*. 81. 4. pp. 491–515.

Bradley, R. and Wagner, C. (2012) Realistic opinion aggregation: Lehrer-Wagner with a finite set of opinion values. *Episteme*. 9. 2. pp. 91–9.

Chambers, C. (2007) An ordinal characterization of the linear opinion pool. *Economic Theory*. 33. 3. pp. 457–474.

[13] We assume that each p_i belongs to $\mathcal{P}'$, i.e., is a regular probability function. So the profile $(p_1, \ldots, p_n)$ is in the domain of the pooling function.

DeGroot, M. H. (1974) Reaching a consensus. *Journal of the American Statistical Association*. 69. 345. pp. 118–121.
Dietrich, F. (2007) A generalised model of judgment aggregation. *Social Choice and Welfare*. 28. 4. pp. 529–65.
Dietrich, F. (2010) Bayesian group belief. *Social Choice and Welfare*. 35. 4. pp. 595–626.
Dietrich, F. and List, C. (2004) A model of jury decisions where all jurors have the same evidence. *Synthese*. 142. 2. pp. 175–202.
Dietrich, F. and List, C. (2007) Arrow's theorem in judgment aggregation. *Social Choice and Welfare*. 29. 1. pp. 19–33.
Dietrich, F. and List, C. (2010) The aggregation of propositional attitudes: Towards a general theory. In Gendler, T. S. and Hawthome, J. (eds.) *Oxford Studies in Epistemology*. Vol. 3. pp. 215–34.
Dietrich, F. and List, C. (2013a) Opinion pooling generalized – Part one: general agendas. Working paper. London School of Economics.
Dietrich, F. and List, C. (2013b) Opinion pooling generalized – Part two: the premise-based approach. Working paper. London School of Economics.
Dietrich, F. and Spiekermann, K. (2013) Independent opinions? On the causal foundations of belief formation and jury theorems. *Mind*. 122. 487. pp. 655–85.
Dokow, E. and Holzman, R. (2010) Aggregation of binary evaluations. *Journal of Economic Theory*. 145. 2. pp. 495–511.
Duddy, C. and Piggins, A. (2013) Many-valued judgment aggregation: Characterizing the possibility/impossibility boundary. *Journal of Economic Theory*. 148. 2. pp. 793–805.
Genest, C. (1984a) A characterization theorem for externally Bayesian groups. *Annals of Statistics*. 12. 3. pp. 1100–5.
Genest, C. (1984b) Pooling operators with the marginalization property. *Canadian Journal of Statistics*. 12. 2. pp. 153–63.
Genest, C., McConway, K. J., and Schervish, M. J. (1986) Characterization of externally Bayesian pooling operators. *Annals of Statistics*. 14. 2. pp. 487–501.
Genest, C. and Wagner, C. (1987) Further evidence against independence preservation in expert judgement synthesis. *Aequationes Mathematicae*. 32. 1. pp. 74–86.
Genest, C. and Zidek, J. V. (1986) Combining probability distributions: A critique and annotated bibliography. *Statistical Science*. 1. 1. pp. 114–35.
Herzberg, F. (2015) Universal algebra for general aggregation theory: Many-valued propositional-attitude aggregators as MV-homomorphisms. *Journal of Logic and Computation*. 25. 3. pp. 965–77.
Lehrer, K. and Wagner, C. (1981) *Rational Consensus in Science and Society*. Dordrecht: Reidel.
List, C. (2012) The theory of judgment aggregation: An introductory review. *Synthese*. 187. 1. pp. 179–207.
List, C. (2013) Social choice theory. *The Stanford Encyclopedia of Philosophy*. Winter [Online] Available from: <http://plato.stanford.edu/archives/win2013/entries/social-choice/> [Accessed: 24 Aug 2015]
List, C. and Pettit, P. (2002) Aggregating sets of judgments: An impossibility result. *Economics and Philosophy*. 18. 1. pp. 89–110.
Madansky, A. (1964) Externally Bayesian groups. Technical Report RM-4141-PR. Santa Monica, CA: RAND Corporation.
McConway, K. J. (1981) Marginalization and linear opinion pools. *Journal of the American Statistical Association*. 76. 374. pp. 410–14.

Mongin, P. (1995) Consistent Bayesian aggregation. *Journal of Economic Theory*. 66. 2. pp. 313–51.
Moral, S. and Sagrado, J. (1998) Aggregation of imprecise probabilities. In Bouchon-Meunier, B. (ed.) *Aggregation and Fusion of Imperfect Information*. pp. 162–88. Heidelberg: Physica-Verlag.
Morris, P. A. (1974) Decision analysis expert use. *Management Science*. 20. 9. pp. 1233–41.
Nehring, K. and Puppe, C. (2010) Abstract Arrovian aggregation. *Journal of Economic Theory*. 145. 2. pp. 467–94.
Pauly, M. and van Hees, M. (2006) Logical constraints on judgment aggregation. *Journal of Philosophical Logic*. 35. pp. 569–585.
Raiffa, H. (1968) *Decision Analysis: Introductory Lectures on Decision Analysis*. Reading, MA: Addison-Wesley.
Spohn, W. (2012) *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.
Stone, M. (1961) The opinion pool. *Annals of Mathematical Statistics*. 32. 4. pp. 1339–42.
Wagner, C. (1982) Allocation, Lehrer models, and the consensus of probabilities. *Theory and Decision*. 14. 2. pp. 207–20.
Wagner, C. (1984) Aggregating subjective probabilities: Some limitative theorems. *Notre Dame Journal of Formal Logic*. 25. 3. pp. 233–40.
Wagner, C. (1985) On the formal properties of weighted averaging as a method of aggregation. *Synthese*. 62. 1. pp. 97–108.
Weymark, J. (1997) Aggregating ordinal probabilities on finite sets. *Journal of Economic Theory*. 75. 2. pp. 407–32.

PART VI

APPLICATIONS OF PROBABILITY: SCIENCE

CHAPTER 26

QUANTUM PROBABILITY

An Introduction

GUIDO BACCIAGALUPPI

[A longer version of this chapter is available online at http://philsci-archive.pitt.edu/10614/. It includes extensive footnotes, further references, and an Appendix with the proofs of the Lemma and Proposition of Section 26.5.]

The topic of probability in quantum mechanics is rather vast, and in this chapter, we choose to discuss it from the perspective of whether and in what sense quantum mechanics requires a generalization of the usual (Kolmogorovian) concept of probability. We shall focus on the case of finite-dimensional quantum mechanics (which is analogous to that of discrete probability spaces), partly for simplicity and partly for ease of generalization. While we shall focus largely on formal aspects of quantum probability (in particular the non-existence of joint distributions for incompatible observables), our discussion will relate also to notorious issues in the interpretation of quantum mechanics. Indeed, whether quantum probability can or cannot be ultimately reduced to classical probability connects rather nicely to the question of 'hidden variables' in quantum mechanics.

26.1 Quantum Mechanics (Once Over Gently)

If a spinning charged object flies through an appropriately inhomogeneous magnetic field, then according to the laws of classical physics it will experience a deflection, which is in the direction of the gradient of the field, proportional to its angular momentum in the same direction (i.e., proportional to how fast it is spinning along that axis). In quantum mechanics one observes a similar effect, except that the object is deflected only by discrete amounts, as if its classical angular momentum could take only certain values (varying by units of Planck's constant $\hbar$). Some particles even seem to possess an intrinsic such 'spin' (i.e., not derived from any rotational motion), e.g. so-called spin-1/2 systems, such as electrons, which get deflected by amounts corresponding to spin values $\pm\frac{\hbar}{2}$. Experiments of this kind are known

as Stern–Gerlach experiments, and they can be used to illustrate some of the most important features of quantum mechanics.

Imagine a beam of electrons, say, moving along the y-axis and encountering a region with a magnetic field inhomogeneous along the x-axis. The beam will split in two, as can be ascertained by placing a screen on the other side of the experiment, and observing that particle detections are localized around two distinct spots on the screen (needless to say, real experiments are a little messier than this).

The first thing to point out is that if we send identically prepared electrons one by one through such an apparatus, each of them will trigger only one detection, *either* in the upper half *or* the lower half of the screen, with probabilities depending on the initial preparation of the incoming electrons.

The second thing to point out is that the same is true whatever the direction in which the inhomogeneous magnetic field is laid, whether along the x-axis, the z-axis, or in any other direction: the beam of electrons will always be split in two components, corresponding to a spin value $\pm\frac{\hbar}{2}$ along the direction of inhomogeneity of the field. If the incoming beam happens to be prepared by selecting one of the deflected beams in a previous Stern–Gerlach experiment (say, a beam of 'spin-x up' electrons), then the probabilities for detection in a further Stern–Gerlach experiment in a direction x' depend only on the angle ϑ between x and x' (and are given by $\cos^2(\vartheta/2)$ and $\sin^2(\vartheta/2)$). So, for example, the probability of measuring spin-z 'up' or 'down' in a beam of spin-x 'up' electrons is 1/2.

For each given preparation procedure (each prepared 'state') we thus have well-defined probability measures over the outcome spaces of various experiments. It is not obvious, however, what in general (if any) should be the *joint* distribution for the outcomes of different experiments, because performing one kind of experiment (say, measuring spin-z) disturbs the probabilities relating to other subsequent experiments (say, spin-x). Indeed, if we imagine performing a spin-z followed by a spin-x measurement on electrons originally prepared in a spin-x up state, we shall get a 50–50 distribution for the results of the last measurement (whether we previously got spin-z up or down), although the original beam was 100% spin-x up. At least in this sense, different measurements in general are incompatible.

What is truly remarkable, however, and makes a straightforward hypothesis of disturbance untenable, is that such a putative disturbance can be *undone* if the spin-z up and spin-z down beam are brought together again before the spin-x measurement. In this case one obtains again spin-x up with probability 1, and this *even* if the whole experiment is performed on one electron at a time. Thus this case cannot be explained by interaction between different electrons when the two beams are brought together again. It reminds one rather of typically wave-like phenomena: the up components of the two beams appear to have interfered constructively, and the down components of the two beams appear to have interfered destructively, although it is each individual electron that displays this wave-like interference behaviour.

There is thus a genuine puzzle (one of many!) about whether and how the probability measures defined over the outcome spaces of the different experiments can be combined. This will be one of the main questions discussed in this chapter.

Let us very briefly review some standard bits of formalism (mainly to ease the transition into the more abstract setting of Section 26.3).[1] The natural mathematical framework for

[1] For more comprehensive but still relatively gentle introductions to quantum mechanics (and its philosophy), see e.g. Albert (1992), Ghirardi (1997), Wallace (2008), Bacciagaluppi (2013), and the

describing interference phenomena is a vector space, where any two elements, call them 'states', $|\psi\rangle$ and $|\varphi\rangle$ can be linearly superposed,

$$\alpha|\psi\rangle + \beta|\varphi\rangle \,. \tag{26.1}$$

The spin degree of freedom of an electron is described by a two-dimensional complex vector space (with the usual scalar product, which we denote $\langle\varphi|\psi\rangle$). Vectors are usually normalized to unit length, since any two vectors $|\psi\rangle$ and $\alpha|\psi\rangle$ that are multiples of each other are considered physically equivalent. Each pair of up and down states is taken to correspond to an orthonormal basis, e.g. the spin-x and spin-z states, related by

$$|+_z\rangle = \frac{1}{\sqrt{2}}(|+_x\rangle + |-_x\rangle)\,, \qquad |-_z\rangle = \frac{1}{\sqrt{2}}(|+_x\rangle - |-_x\rangle). \tag{26.2}$$

Thus, while each spin-z state can be split into both up and down components in the spin-x basis, the down components, say, can cancel out again if the two spin-z states are appropriately combined:

$$\frac{1}{\sqrt{2}}(|+_z\rangle + |-_z\rangle) = \frac{1}{\sqrt{2}}\Big(\frac{1}{\sqrt{2}}(|+_x\rangle + |-_x\rangle) + \frac{1}{\sqrt{2}}(|+_x\rangle - |-_x\rangle)\Big) = |+_x\rangle \,. \tag{26.3}$$

Temporal evolution is given by the action of an appropriate group of linear operators (i.e., of linear mappings, which map superpositions into the corresponding superpositions) on the states:

$$|\psi\rangle \mapsto U_{t_2 t_1}|\psi\rangle \,. \tag{26.4}$$

These operators are in fact unitary, i.e., preserve the length of vectors (and more generally scalar products between them). In a Stern–Gerlach measurement, the relevant unitary evolution is generated by an operator containing a term proportional to a 'spin operator', e.g. the z-spin operator S_z, written

$$S_z = \frac{\hbar}{2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \tag{26.5}$$

in the spin-z basis, which thus simply multiplies by a scalar the spin-z vectors:

$$S_z|\pm_z\rangle = \pm\frac{\hbar}{2}|\pm_z\rangle \tag{26.6}$$

(the spin-z states are *eigenstates* of the operator with corresponding *eigenvalues* $\pm\frac{\hbar}{2}$). During the measurement, this term couples the spin-z eigenstates (which it leaves invariant, perhaps up to a complex phase) to the spatial degrees of freedom of the electron, thus deflecting the motion of the electron:

$$|\pm_z\rangle|\psi\rangle \mapsto U|\pm_z\rangle|\psi\rangle = |\pm_z\rangle|\psi_\pm\rangle \tag{26.7}$$

relevant articles in the *Stanford Encyclopedia of Philosophy* (http://plato.stanford.edu/). For treatments emphasizing the modern notion of effect-valued observables, used below in Section 26.3, see Busch, Lahti, and Mittelstaedt (1991) and Busch, Grabowski, and Lahti (1995).

(where we assume that $|\psi_\pm\rangle$ are states of the spatial degrees of freedom of the electron in which the electron is localized in two non-overlapping regions).

Measurable quantities in quantum mechanics (usually called 'observables') are traditionally associated with such operators, the corresponding eigenvalues being the values the observable can take.[2] Two observables thus understood will be compatible if they have all eigenvectors in common, or equivalently if the associated operators commute, i.e., $AB|\psi\rangle = BA|\psi\rangle$ for all states $|\psi\rangle$. Incompatibility of quantum mechanical observables is intuitively related to the idea that measurements of non-commuting observables generally require *mutually exclusive* experimental arrangements (implemented through appropriate unitary operators).

Note that if the initial state of the electron is a superposition of spin-z states, e.g., the spin-x up state (26.3), the linearity of the evolution will preserve the superposition:

$$\frac{1}{\sqrt{2}}(|+_z\rangle + |-_z\rangle)|\psi\rangle \mapsto \frac{1}{\sqrt{2}}(|+_z\rangle|\psi_+\rangle + |-_z\rangle|\psi_-\rangle)\,. \tag{26.8}$$

State (26.8), unlike states (26.7) is no longer a product state. Indeed, in quantum mechanics the composition of degrees of freedom (or of different systems) is described using the *tensor product* of the vector spaces describing the different degrees of freedom (or systems), which is the space of all linear superpositions of product states. Non-product states are called *entangled*, and are the source of some of the most peculiar features of quantum mechanics. We can ignore them further, however, until we make contact with the discussion of the Bell inequalities in Section 26.5.

The last thing we need to recall from elementary treatments of quantum mechanics is what happens upon measurement. Namely, when a certain observable is measured, say S_z, the state of the system undergoes a stochastic transformation (a 'collapse', or 'reduction', or 'projection'), given mathematically by the projection on to one of the eigenstates of the measured observable, i.e., by the application of the projection operator $P^z_+ = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ (or $P^z_- = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$) — the operator with eigenvectors $|\pm_z\rangle$ and corresponding eigenvalues 1 and 0 (or 0 and 1). Thus, for instance,

$$|+_x\rangle = \frac{1}{\sqrt{2}}(|+_z\rangle + |-_z\rangle) \mapsto P^z_+ \frac{1}{\sqrt{2}}(|+_z\rangle + |-_z\rangle) = \frac{1}{\sqrt{2}}|+_z\rangle \tag{26.9}$$

or

$$|+_x\rangle \mapsto P^z_- \frac{1}{\sqrt{2}}(|+_z\rangle + |-_z\rangle) = \frac{1}{\sqrt{2}}|-_z\rangle\,. \tag{26.10}$$

(The final state is then thought of as renormalized, i.e., rescaled to the unit vector $|+_z\rangle$ or $|-_z\rangle$, respectively.)

[2] In order for such an operator to generate a unitary group, it has to be self-adjoint, which in finite dimensions means simply that the corresponding matrix is conjugate symmetric, e.g., $\begin{pmatrix} \alpha^* & \gamma^* \\ \beta^* & \delta^* \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$. This implies further that all its eigenvalues are real and that the vector space has an orthonormal basis composed of eigenvectors of the operator.

The probability for such a transformation is given by the so-called 'Born rule': it is the modulus squared of the coefficient of the corresponding component in the initial state, or equivalently the squared norm of the (unnormalized) collapsed state. In the cases of both (26.9) and (26.10) this equals 1/2. (Note that in a Stern–Gerlach experiment what we measure is in fact whether the electron impinges on the upper or the lower half of the screen, thus collapsing the state (26.8) to one of $\frac{1}{\sqrt{2}}|\pm_z\rangle|\psi_\pm\rangle$, each with probability 1/2.)

26.2 Classical Probability (with an Eye to Quantum Mechanics)

We shall now look at the usual Kolmogorovian notion of probability, formulated with an emphasis on aspects relevant for the analogy with quantum probability in the next section. To this end, we introduce a (suitably general) notion of *observable*. We consider only probability spaces with discrete (finite or denumerable) event spaces Ω, and assume that the measurable sets $\mathcal{B}$ include all the singletons $\{\omega\} \subset \Omega$. Consider first the random variables with values in the unit interval $e : \Omega \to [0, 1]$ (so-called *response functions* or *effects*), and as a special case the response functions $\chi : \Omega \to \{0, 1\}$ (these are identical with the *characteristic functions* of measurable subsets $\Sigma \subset \Omega$, i.e. functions that take the value 1 on Σ and 0 on $\Omega \setminus \Sigma$). We now define an *observable* as a (finite or denumerable) family of response functions $\{e_i\}_{i\in I}$ such that

$$\sum_{i\in I} e_i = 1 \tag{26.11}$$

(where 1 is the random variable that is identically 1).

For each such observable, the probability measure p induces a probability measure on (the Boolean algebra generated by) the family of functions $\{e_i\}$ (or on the index set I), which we also denote by p:

$$p(e_i) := \sum_{\omega\in\Omega} e_i(\omega)p(\{\omega\}) \,, \tag{26.12}$$

and for any subset J of I:

$$p(\sum_{i\in J} e_i) := \sum_{i\in J}\sum_{\omega\in\Omega} e_i(\omega)p(\{\omega\}) \,. \tag{26.13}$$

(This is correctly normalized because of (26.11) and additivity.) In the special case in which all e_i are 'sharp' ($e_i(1 - e_i) = 0$, where 0 is the random variable that is identically 0) — i.e., characteristic functions —, we see that the probabilities are just the measures of the sets defined by the characteristic functions $\sum_{i\in J} e_i$, so that the 'sharp observables' are in bijective correspondence with the (finite or denumerable) partitions of Ω, and 'measuring' a sharp observable is simply a procedure for distinguishing between the elements of such a partition.

General observables (at least in the classical case) can be interpreted as noisy or fuzzy or unsharp versions of sharp observables. Indeed, take the following observable, given by the

resolution of the identity

$$\sum_{\omega\in\Omega}\chi_{\{\omega\}}=1 \tag{26.14}$$

(we can call this the *finest sharp observable*). Now, every effect e can be written as

$$\sum_{\omega\in\Omega}e(\omega)\chi_{\{\omega\}}\ . \tag{26.15}$$

Since for any observable $\{e_i\}$, in particular also for (26.14), the probability of each e_i is given by (26.12), we have that

$$p(\chi_{\{\omega\}})=\sum_{\omega'\in\Omega}\chi_{\{\omega\}}p(\{\omega'\})=\chi_{\{\omega\}}p(\{\omega\})=p(\{\omega\})\ , \tag{26.16}$$

and thus

$$p(e_i)=\sum_{\omega\in\Omega}e_i(\omega)p(\chi_{\{\omega\}})\ . \tag{26.17}$$

And we see that $e_i(\omega)$ can be interpreted as the *conditional probability* for the response e_i in the experiment $\{e_i\}$, given that a (counterfactual) measurement of the finest sharp observable would have yielded ω.

It is important to note that while each experiment has a Boolean structure (the subsets of the index set I form a Boolean algebra), in general these Boolean algebras do *not* correspond to the Boolean subalgebras of $\mathcal{B}$. It is only the Boolean algebras associated with measurements of sharp observables that correspond to Boolean subalgebras of $\mathcal{B}$.

There are two further notions we wish to introduce with an eye to the analogy with quantum probability. One is a notion of *compatibility* of observables. To this end, we first introduce the *coarse-graining* of observables: the observable $\{e_i\}_{i\in I}$ is a coarse-graining of the observable $\{g_k\}_{k\in K}$ iff there is a partition of the index set $K=\cup_{i\in I}K_i$ such that for all $i\in I$,

$$e_i=\sum_{k\in K_i}g_k \tag{26.18}$$

(note that every sharp observable is a coarse-graining of the finest sharp observable (26.14)). Clearly, any experiment that measures $\{g_k\}$ also measures $\{e_i\}$. We now call two observables $\{e_i\}$ and $\{f_j\}$ *compatible* iff there is an observable $\{g_k\}$ such that $\{e_i\}$ and $\{f_j\}$ are both coarse-grainings of $\{g_k\}$. The observable $\{g_k\}$ is called a *joint observable* for (or a *joint fine-graining* of) $\{e_i\}$ and $\{f_j\}$.

Obviously, *any two classical observables are compatible*. Indeed, given any two observables $\{e_i\}$ and $\{f_j\}$, we can define a joint observable $\{g_k\}$ simply as $\{e_if_j\}_{(i,j)}$ (the indices ranging over those pairs (i,j) for which the product $e_if_j\neq 0$). In the special case of sharp observables, the product e_if_j is of course the characteristic function of the intersection of the sets defined by e_i and f_j, and the joint observable corresponds simply to the Boolean subalgebra of $\mathcal{B}$ generated by the union of the two subalgebras corresponding to $\{e_i\}$ and $\{f_j\}$.

The other notion we introduce with an eye to quantum probability is that of a *state*, defined as a *family of overlapping probability measures* over the outcomes of all possible experiments, i.e., a mapping p from the response functions to the reals, such that:

(C1) For all e, $p(e) \in [0,1]$,
(C2) $p(1) = 1$,
(C3) For all (finite or denumerable) families $\{e_i\}$ of effects with $\sum_i e_i \leq 1$,

$$p(\sum_i e_i) = \sum_i p(e_i) \ .$$

We have already seen that a classical probability measure p induces a probability measure on every classical observable and thus defines a state in this sense. Conversely, the family of the probability measures on all observables fixes the original probability measure uniquely, because the original probability measure is nothing other than the probability measure associated with the finest sharp observable (26.14).

26.3 Quantum Mechanics (with an Eye to Probability)

As mentioned in Section 26.1, the formalism of quantum mechanics is based on the fairly familiar structure of a vector space with scalar product, or technically a Hilbert space (because it is complete in the norm induced by the scalar product). What interests us in particular (with an eye to highlighting the probabilistic structure of quantum mechanics) are the notions of 'state' and 'observable' we find in the formalism.

States are associated in the first place with (unit) vectors in the space. As in our example of a Stern–Gerlach measurement and the associated 'collapse' of the state (26.9–26.10), experiments are generally and abstractly associated with probabilistic transformations of the states, corresponding to the idea that 'measurements' induce an irreducible disturbance of a quantum system. Restricting ourselves to unit vectors, such transformations can be written in general as:

$$|\psi\rangle \mapsto A|\psi\rangle \tag{26.19}$$

(where the right-hand side should be thought of as suitably renormalized). The probability for a transition of the form (26.19) is given by

$$\langle\psi|A^*A|\psi\rangle \ , \tag{26.20}$$

which is the scalar product of the vector $A|\psi\rangle$ with itself.[3] The operator A in (26.19) is arbitrary (in particular not necessarily unitary or self-adjoint), the only restriction being that (26.20) be no greater than 1. Note that the probability (26.20) depends only on the product A^*A and not on the specific transformation A (indeed, one can construct infinitely many other operators B such that $B^*B = A^*A$). Operators of the form $E = A^*A$ for some transformation A are called 'effects'.

[3] The operator A^*, called the adjoint of A, is the operator defined by $\langle\varphi|A^*\psi\rangle = \langle A\varphi|\psi\rangle$. Self-adjoint operators are operators with $A^* = A$. Note that $(AB)^* = B^*A^*$, so that in particular an operator of the form A^*A is self-adjoint. Note also that for A self-adjoint, the mapping $A \mapsto \langle\psi|A\psi\rangle$ (normally written $A \mapsto \langle\psi|A|\psi\rangle$) is a linear functional on to the positive reals.

Each experiment will include a number of alternative transformations that could possibly take place:

$$|\psi\rangle \mapsto A_i|\psi\rangle \,, \tag{26.21}$$

whose combined probability equals 1:

$$\sum_i \langle\psi|A_i^* A_i|\psi\rangle = 1 \,. \tag{26.22}$$

This is required to hold for all possible unit vectors $|\psi\rangle$, so in fact we have

$$\sum_i A_i^* A_i = \mathbf{1} \,, \tag{26.23}$$

or, writing $E_i := A_i^* A_i$,

$$\sum_i E_i = \mathbf{1} \,, \tag{26.24}$$

where $\mathbf{1}$ is the identity operator. Each such (finite or denumerable) family of effects $\{E_i\}_{i\in I}$, or 'resolution of the identity' (26.24) is called an *observable*, quantum mechanical effects being the formal analogue of classical response functions.

Note that the probability of an effect (in any given state) is independent of which observable the effect is part of, i.e., which family of alternative transformations is being implemented in a particular experiment. We shall return to this 'non-contextuality' of probabilities in Sections 26.5 and 26.6. Suffice it to say now that it is a non-trivial feature because, unlike the classical case, the same effect could be part of two mutually *incompatible* observables.

There is more than one definition of (in)compatibility in the literature, but the following one (on which we have modelled the definition of Section 26.2) is the most suited to our purposes (see e.g. Cattaneo *et al.* 1997). Define an observable $\{E_i\}_{i\in I}$ to be a *coarse-graining* of the observable $\{G_k\}_{k\in K}$ iff there is a partition of the index set $K = \cup_{i\in I} K_i$ such that for all $i \in I$,

$$E_i = \sum_{k\in K_i} G_k \,. \tag{26.25}$$

Any experiment that measures $\{G_k\}$ also measures $\{E_i\}$. As in the classical case, we call two observables $\{E_i\}$ and $\{F_j\}$ *compatible* iff there is an observable $\{G_k\}$ such that $\{E_i\}$ and $\{F_j\}$ are both coarse-grainings of $\{G_k\}$. The observable $\{G_k\}$ is called a *joint observable* for (or a *joint fine-graining* of) $\{E_i\}$ and $\{F_j\}$. Compatibility of two observables can be easily generalized to joint compatibility of arbitrary sets of observables.

The definition of an observable in any (older) textbook on quantum mechanics is as a self-adjoint operator A, i.e., an operator with $A^* = A$. But this traditional definition corresponds to a special case of the one above. Self-adjoint operators are diagonalizable, in the sense that $A = \sum_i a_i P_i$ with real a_i (the eigenvalues) and $\{P_i\}$ a family of projections (self-adjoint operators with $P_i^2 = P_i$, or $P_i(\mathbf{1} - P_i) = \mathbf{0}$, where $\mathbf{0}$ is the zero operator) that are mutually orthogonal ($P_i P_j = \mathbf{0}$ for $i \neq j$). Thus, each self-adjoint operator is associated with a unique 'projection-valued observable', i.e., a resolution of the identity (26.24), in which the effects E_i are in fact projections (they are 'sharp', meaning $E_i(\mathbf{1} - E_i) = \mathbf{0}$), and which is finite if the Hilbert space is finite-dimensional. Note also that a measurement of such a

'sharp observable' can be implemented by taking $A_i = P_i$, since $P_i^* P_i = P_i$, i.e., each state is transformed to an eigenstate of the measured observable. This is the usual 'collapse postulate' or 'projection postulate' of textbook quantum mechanics, corresponding to a 'minimally disturbing' measurement of a sharp observable.

Compatibility of two sharp observables A and B is traditionally defined as their *commutativity*, i.e., $AB = BA$. This is equivalent to the commutativity of the elements of the respective resolutions of the identity, $\sum_i P_i = \mathbf{1}$ and $\sum_j Q_j = \mathbf{1}$, i.e.,

$$P_i Q_j = Q_j P_i \qquad \text{for all } i,j\,. \tag{26.26}$$

In this case, a joint (projection-valued) observable $\{R_k\}$ is given by $\{P_i Q_j\}_{(i,j)}$ (the indices (i,j) ranging over those pairs for which $P_i Q_j \neq 0$). Indeed, trivially,

$$P_i = \sum_n P_i Q_n \qquad \text{and} \qquad Q_j = \sum_m P_m Q_j\,. \tag{26.27}$$

This argument generalizes to show that finite sets of pairwise compatible sharp observables possess a joint projection-valued resolution of the identity. Indeed, since in finite dimensions all diagonal decompositions are discrete, one can generalize it further to arbitrary sets of pairwise commuting operators.

Effect-valued and projection-valued observables are the analogues, respectively, of the general and sharp classical observables of Section 26.2, and, at least in some cases, also effect-valued observables can be interpreted as 'unsharp' versions of projection-valued ones. Since every effect E is a self-adjoint operator, it is itself diagonalizable as

$$E = \sum_k e_k P_k\,, \tag{26.28}$$

where all the eigenvalues e_k are positive and lie in the interval $[0,1]$. Now suppose all effects E_i in an observable (26.24) *commute*: there will then exist a single projection-valued resolution of the identity $\{R_k\}$, such that every E_i can be written as

$$E_i = \sum_k e_k^i R_k\,. \tag{26.29}$$

with suitable coefficients e_k^i. The probability of E_i in the state $|\psi\rangle$ is

$$\langle\psi|E_i|\psi\rangle = \sum_k e_k^i \langle\psi|R_k|\psi\rangle\,, \tag{26.30}$$

and we can again at least formally identify the coefficient e_k^i as the *conditional probability* that the measurement of $\{E_i\}$ yields i given that a measurement of $\{R_k\}$ would have yielded k (it is in fact correctly normalized). Thus, we can think of a commutative effect-valued observable as (at least probabilistically equivalent to) a 'noisy' or 'fuzzy' or 'unsharp' measurement of an associated projection-valued observable.

Further, given our definition of joint observables, we can also understand the general case of a *non-commutative* effect-valued observable $\{E_i\}_{i\in I}$ as a *joint observable* for the generally denumerably many *commutative* effect-valued observables $\{E_i, \mathbf{1} - E_i\}$, one for

each $i \in I$. This gives us a further insight on compatibility and incompatibility — namely that incompatible observables can be made compatible if one is willing to introduce enough 'noise' in one's measurements.

While in a sense we can thus reduce the effect-valued observables to the projection-valued ones (and the question of incompatibility essentially to that of incompatibility for projection-valued observables), it makes very good sense to work with the more general effect-valued ones. To quote three reasons: effect-valued observables are needed for modelling realistic experiments; the concatenation of two experiments is clearly an experiment, but cannot generally be represented by a projection-valued observable; and no measurement of a projection-valued observable can fully determine a quantum state, while — precisely because effect-valued observables can combine probabilistic information from incompatible projection-valued ones, even if noisily — there are so-called 'informationally complete' effect-valued observables, whose measurement statistics completely determine the quantum state.

Let us now return to the quantum states themselves. We can think of a quantum state as defining a *family of overlapping probability measures* over the outcomes of all possible experiments. More precisely, we can identify states $|\psi\rangle$ with mappings p_ψ from the effects to the reals, such that:

(Q1) For all E, $p_\psi(E) = \langle\psi|E|\psi\rangle \in [0,1]$,
(Q2) $p_\psi(\mathbf{1}) = \langle\psi|\mathbf{1}|\psi\rangle = 1$,
(Q3) For all (finite or denumerable) families $\{E_i\}$ of effects with $\sum_i E_i \leq \mathbf{1}$,

$$p_\psi(\sum_i E_i) = \sum_i p_\psi(E_i) \, .$$

This definition can of course be restricted to projection operators (with (Q3) restricted to families of mutually orthogonal projections). If one does so, a famous theorem due to Gleason (1957/1975) (valid for Hilbert spaces of dimension $n \geq 3$) shows that the most general state ρ on the projections is an *arbitrary convex combination* (i.e., a weighted average) of vector states p_ψ. *A fortiori* this is true for states defined on effects (and the direct proof for this case is much simpler (Busch 2003)). The most general quantum states thus form a convex set, with the vector states as its extremal points.

Perhaps surprisingly, these general quantum states cannot *uniquely* be decomposed as convex combinations of vector states. (We shall not have the space to develop this point, but it is a very important disanalogy with the classical case.) This can be seen very easily in the special case of a spin-1/2 system (described by a 2-dimensional Hilbert space), using a rather beautiful geometric representation, the so-called Poincaré sphere (or Bloch sphere). The unit vectors form a 2-dimensional complex sphere, and this is affinely isomorphic to a 3-dimensional real sphere. That is, one can map the two bijectively in a way that preserves convex combinations. This allows one to associate the abstract 'spin' states with directions in 3-dimensional space: an electron has spin 'up' in the direction r iff its abstract spin state is mapped to the spatial vector r under this affine mapping, and 'down' iff it is mapped to the spatial vector $-r$. One can then see directly that any point in the interior of the sphere (corresponding to a general quantum state) can be written in infinitely many ways as a convex combination of points on the surface of the sphere (corresponding to vector

states). Indeed, any straight line through a point in the interior will intersect the surface of the sphere in two points, thus defining a convex decomposition of the interior point, but there are infinitely many such straight lines. The centre of the sphere corresponds to the so-called maximally mixed state, which assigns equal probabilities 1/2 to spin up or down in any spatial direction *r*.

We are now ready to sketch a generalization of the notion of probability encompassing both classical and quantum probability (Section 26.4), and to discuss whether it is indeed a non-trivial generalization of the classical notion (Sections 26.5 and 26.6).

26.4 Generalized Probability (a Sketch)

There are different (but largely convergent) approaches to defining a theory of probability generalizing both classical and quantum probability. The seminal work in this area is due to Mackey (1957), who inaugurated what is known today as the 'convex set' approach to quantum and generalized probability, which axiomatizes directly pairs of state-observable structures (Mackey 1963, Varadarajan 1968, Beltrametti and Cassinelli 1981).

Alternative more concrete routes to generalized probability have been pursued in particular by Ludwig (1954, 1985) and his group, and by Foulis and Randall with their work on test spaces (starting with Randall and Foulis (1970, 1973), and Foulis and Randall (1972, 1974); see also the review by Wilce (2000)).

Research in quantum logic opened yet further routes into generalized probability by providing various generalizations of Boolean algebras (orthomodular lattices, orthomodular posets, partial Boolean algebras, orthoalgebras, effect algebras etc.), thus generalizing the event spaces of classical probability. The main lines of research in this tradition stem, respectively, from the classic paper by Birkhoff and von Neumann (1936/1975), which inaugurated the lattice-theoretic version of quantum logic (with its emphasis on weakening the distributive law, originally in favour of modularity and then of orthomodularity), and from the work on partial Boolean algebras by Specker (1960) and Kochen and Specker (1965/1975a,b, 1967/1975) (with its emphasis on partial operations).

I shall deliberately ignore these distinctions and sketch instead a somewhat pedagogical version of generalized probability theory — drawing from elements of these various approaches, and having the advantage of being fairly simple and of leading rather naturally to the abstract notions of effect algebras and orthoalgebras (possibly the best current candidates for providing an abstract setting for a generalized probability theory).[4] While much of what follows can be generalized or ought to be generalizable to the denumerable case, in this section I shall focus exclusively on the finite case.

Let us start with the quasi-operational idea of a set $\mathcal{A}$ of *experiments*. Each experiment $A \in \mathcal{A}$ is characterized by a Boolean algebra $\mathcal{B}_A$ of experimental outcomes $e_A \in \mathcal{B}_A$. We further consider *states* $p \in \mathcal{P}$, which define probability measures over the outcomes of all

[4] Two other approaches that unfortunately I shall have to ignore are the generalizations of probability theory based on upper and lower probabilities, and on negative probabilities; see e.g. de Barros and Suppes (2010) for the former, and Abramsky and Brandenburger (2011) and Oas, de Barros, and Carvalhaes (2014) for the latter (and references therein).

possible experiments, $p(e_A)$. If an outcome e_A of one experiment can be somehow identified with an outcome e_B of a different experiment, one will obviously require that $p(e_A) = p(e_B)$ for all states p, i.e., the probability measures induced by the states can generally overlap. Less obviously, we shall identify *all* pairs of experimental outcomes that are equiprobable in all states. This will be the first of only very few substantial requirements. Thus we shall define *effects* as equivalence classes $e = [e_A]$ of experimental outcomes under the equivalence relation $\sim$ of equiprobability for all p:

$$e_A \sim e_B \quad :\Leftrightarrow \quad p(e_A) = p(e_B) \text{ for all } p\,. \tag{26.31}$$

We denote the set of all effects by $\mathcal{E}$. In Section 26.6 we shall return to the question of identifying outcomes of different experiments. Suffice it to say at this stage that identifying equiprobable outcomes means we are thinking of effects as characterized by what they can tell us about the states. Note that each effect e defines an affine mapping from the states to the unit interval:

$$e : \mathcal{P} \to [0,1], \quad p \mapsto p(e)\,. \tag{26.32}$$

It may be convenient to require that every such mapping corresponds to an effect, but for our limited purposes we shall not do so.

Now, for any two experiments $A, B \in \mathcal{A}$,

$$0_A \sim 0_B \quad \text{and} \quad 1_A \sim 1_B \tag{26.33}$$

(where 0_A and 1_A are the 0 and 1 elements of the Boolean algebra $\mathcal{B}_A$, and similarly for $\mathcal{B}_B$). Indeed, $p(0_A) = 0$ and $p(1_A) = 1$ for all p, independently of A. We can thus define an effect 0 and an effect 1 as

$$0 := [0_A] \quad \text{independently of } A \tag{26.34}$$

and

$$1 := [1_A] \quad \text{independently of } A\,. \tag{26.35}$$

Similarly, for any $A, B \in \mathcal{A}$, if $e_A \sim e_B$ then $\neg e_A \sim \neg e_B$ (where $\neg$ denotes negation in the relevant Boolean algebra). Indeed, if $p(e_A) = p(e_B)$ for all p, then

$$p(\neg e_A) = 1 - p(e_A) = 1 - p(e_B) = p(\neg e_B) \quad \text{for all } p\,, \tag{26.36}$$

and for any effect e we can define a unique effect $e^\perp$ as

$$e^\perp := [\neg e_A] \quad \text{independently of } A\,. \tag{26.37}$$

Clearly, for any e, we have $e^{\perp\perp} = e$. Note, however, that it is perfectly possible for some e that $e^\perp = e$ (this is the case for both the response function $\frac{1}{2}1$ in classical probability and the effect $\frac{1}{2}\mathbf{1}$ in quantum probability).

The states naturally induce a *partial ordering* on the effects — which will be a useful tool in the following — defined as

$$e \leq f \quad :\Leftrightarrow \quad p(e) \leq p(f) \text{ for all } p\,. \tag{26.38}$$

Note that

$$e \leq f \Leftrightarrow f^\perp \leq e^\perp\,. \tag{26.39}$$

We now introduce two important notions: compatibility and orthogonality of effects.

Two effects e and f are *compatible*, written $e\$f$, iff there is an experiment A and outcomes $e_A \in \mathcal{B}_A$ and $f_A \in \mathcal{B}_A$ such that $e = [e_A]$ and $f = [f_A]$. That is, two effects are compatible iff they can be measured in a single experiment. The definition of compatibility can be trivially extended to finite sets of effects.

Two effects e and f are *orthogonal* (or *disjoint*), written $e \perp f$, iff $e \leq f^\perp$, i.e., $p(e) \leq 1 - p(f)$ for all p. Note that this relation is symmetric, but generally not irreflexive (since it is possible that $e^\perp = e$). We can generalize also orthogonality to finite sets of effects, by defining a family $\{e_i\}$ of effects to be jointly orthogonal iff $\sum_i p(e_i) \leq 1$ for all p.

Note that if there are experimental outcomes e_A and f_A in some $A \in \mathcal{A}$ such that $e_A \leq f_A$ with respect to the partial order on the Boolean algebra $\mathcal{B}_A$, then $p(e_A) \leq p(f_A)$ for all p, and therefore $[e_A] \leq [f_A]$ with respect to the partial order on the effects.

We shall *require*, conversely, that if $e \leq f$ for two effects e and f, then there exist at least one experiment A and experimental outcomes $e_A \in e$ and $f_A \in f$, such that $e_A \leq f_A$ in the Boolean algebra $\mathcal{B}_A$. This is the second of our substantive requirements.

It follows in particular that *comparable effects are compatible*, that is, that $e \leq f$ implies that e and f are in fact compatible. Since $e_A \leq f_A$ for some A means that $p(f_A \wedge \neg e_A) = p(f_A) - p(e_A)$ for all p, it also follows that if $e \leq f$, there is an effect $f \ominus e := [f_A \wedge \neg e_A]$ (independently of any particular A with $e_A \leq f_A$), which is jointly compatible with e and f, and such that $p(f \ominus e) = p(f) - p(e)$ for all p.

By this second requirement we also have that *orthogonal effects are compatible*. Indeed, $e \perp f$ means that e and $f^\perp$ are comparable, and thus compatible. But if $e_A \in e$ and $\neg f_A \in f^\perp$ are in the same Boolean algebra $\mathcal{B}_A$, then so are e_A and f_A, thus e and f are compatible. It also follows that if $e \perp f$, i.e., $p(e) + p(f) \leq 1$ for all p, there is an effect $e \oplus f$ jointly compatible with e and f, such that $p(e \oplus f) = p(e) + p(f)$ for all p. Indeed, given the above, we can define

$$e \oplus f := (f^\perp \ominus e)^\perp = [e_A \vee f_A] \tag{26.40}$$

(independently of any particular A with $e_A \leq f_A$), which is jointly compatible with e and f, and such that

$$p(e \oplus f) = p(e) + p(f) \quad \text{for all } p\,. \tag{26.41}$$

Note that for $e \leq f$ we thus have

$$f = e \oplus (f \ominus e)\,, \tag{26.42}$$

or

$$f = e \oplus (e \oplus f^\perp)^\perp \tag{26.43}$$

(the so-called 'effect algebra orthomodular identity').

As our third and last substantive requirement, we shall strengthen the second requirement such that for (finite) *ordered chains* of effects $e^1 \leq e^2 \leq e^3 \leq \ldots$, there shall exist an experiment A and experimental outcomes $e^1_A \in e^1$, $e^2_A \in e^2$, etc., such that $e^1_A \leq e^2_A \leq e^3_A \leq \ldots$ in the Boolean algebra $\mathcal{B}_A$, so that in particular the effects in the chain are jointly compatible. (In the rest of this section, whenever we write 'ordered chain' we shall mean 'finite ordered chain', and similarly for 'jointly orthogonal set'.)

If ordered chains are jointly compatible it follows that also *jointly orthogonal sets* of effects are jointly compatible. Indeed, given a jointly orthogonal set of effects $\{e^i\}_{i=1}^N$, the sequence

of effects

$$e^1,\ e^1 \oplus e^2,\ (e^1 \oplus e^2) \oplus e^3,\ \ldots \tag{26.44}$$

is an ordered chain, and so is a jointly compatible set. But if $e^1_A, e^1_A \vee e^2_A, \ldots$ are in the same Boolean algebra $\mathcal{B}_A$, so are $e^1_A, e^2_A, \ldots$, and the original set $\{e^i\}$ is jointly compatible.

A (finite) observable on $\mathcal{E}$ can now be defined simply as a jointly orthogonal set (for which one automatically has $\bigoplus_i e_i \leq 1$), and a state $p \in \mathcal{P}$ can be identified with a mapping from the effects to the reals, such that:

(G1) For all $e \in \mathcal{E}$, $p(e) \in [0,1]$,
(G2) $p(1) = 1$,
(G3) For all jointly orthogonal sets $\{e_i\}$ of effects,

$$p(\bigoplus_i e_i) = \sum_i p(e_i) \ .$$

Since jointly orthogonal sets are jointly compatible, to each such observable on $\mathcal{E}$ there corresponds at least one experiment $A \in \mathcal{A}$. Coarse-graining and compatibility of observables can be defined as above, and compatibility of two effects e and f is trivially equivalent to compatibility of the two observables $\{e, 1-e\}$ and $\{f, 1-f\}$.

We are now in a position to show that the structure $(\mathcal{E}, 0, 1, \oplus)$ is an *effect algebra*, that is, a structure with two distinguished elements 0 and 1 and a partial operation $\oplus$ (defined on a subset of $\mathcal{E} \times \mathcal{E}$), satisfying the following axioms:

(E1) The partial operation $\oplus$ is commutative, i.e., if $e \oplus f$ is defined, so is $f \oplus e$, and $e \oplus f = f \oplus e$.
(E2) The partial operation $\oplus$ is associative, i.e., if $e \oplus f$ and $(e \oplus f) \oplus g$ are defined, so are $f \oplus g$ and $e \oplus (f \oplus g)$, and $(e \oplus f) \oplus g = e \oplus (f \oplus g)$.
(E3) For any e, there is a unique element $e^\perp$ such that $e \oplus e^\perp = 1$.
(E4) If $e \oplus 1$ is defined, then $e = 0$.

(The elements of an abstract effect algebra are also called effects.)
Proof:
Define $\oplus$ as above. Since the relation $\perp$ is symmetric, if $e \oplus f$ is defined, so is $f \oplus e$, and (E1) follows because

$$p(e) + p(f) = p(f) + p(e) \quad \text{for all } p \ . \tag{26.45}$$

Next, assume that $e \oplus f$ and $(e \oplus f) \oplus g$ are defined, i.e.,

$$p(e) \leq 1 - p(f) \quad \text{and} \quad p(e) + p(f) \leq 1 - p(g) \quad \text{for all } p \ . \tag{26.46}$$

Then also

$$p(f) \leq 1 - p(g) \quad \text{and} \quad p(f) + p(g) \leq 1 - p(e) \quad \text{for all } p \ , \tag{26.47}$$

i.e., also $f \oplus g$ and $e \oplus (f \oplus g)$ are defined, and (E2) follows because

$$\Big(p(e) + p(f)\Big) + p(g) = p(e) + \Big(p(f) + p(g)\Big) \ . \tag{26.48}$$

(Associativity of $\oplus$ was already implicit when we showed above that orthogonal sets are jointly compatible, rather than only orthogonal sequences.)
Further, for each e the unique element satisfying (E3) is the element $e^\perp$ defined by (26.37): clearly $e \oplus e^\perp$ is defined and $e \oplus e^\perp = 1$; and because of (26.41), if there are two effects f and f' both satisfying

$$p(e \oplus f) = p(e \oplus f') = 1 \qquad \text{for all } p\,, \tag{26.49}$$

then $p(f) = p(f')$ for all p, hence $f = f'$, and (E3) follows.
Finally, if $e \oplus 1$ is defined then $p(e \oplus 1) = p(e) + p(1) = p(e) + 1$ for all p, but since $0 \leq p(e), p(e \oplus 1) \leq 1$, we have $p(e) = 0$ for all p, and (E4) follows. *QED*.

Note that in any effect algebra, we can abstractly define a partial order $e \leq f$ as: there is a g such that $e \oplus g = f$. Given our definition of $\oplus$, it follows from (26.41) and (26.42) above that our previous definition of the partial order on $\mathcal{E}$ coincides with the abstract one.

Similarly, in any effect algebra one can abstractly define relations of orthogonality and compatibility. Two effects e and f are orthogonal in the abstract sense iff $e \oplus f$ is defined, and a finite set of effects $\{e_i\}$ is jointly orthogonal iff $\bigoplus_i e_i$ is defined. Our definition of orthogonality for $\mathcal{E}$ clearly coincides with the abstract one.

As for compatibility, two or finitely many effects $\{f_j\}$ are compatible in the abstract sense iff there is an orthogonal set $\{e_i\}_{i \in I}$ and subsets $I_j \subset I$ such that $f_j = \bigoplus_{i \in I_j} e_i$ for all j (we shall say the family of effects $\{f_j\}$ has an orthogonal decomposition).

It is easy to see that also our definition of compatibility for $\mathcal{E}$ coincides with the abstract one. For instance (and similarly for finitely many effects), if two effects e and f are compatible in our sense above, there is an experiment $A \in \mathcal{A}$ and experimental outcomes $e_A \in e$ and $f_A \in f$ in $\mathcal{B}_A$. In this case the effects $g := [e_A \wedge f_A]$, $h := [e_A \wedge \neg f_A]$ and $i := [\neg e_A \wedge f_A]$ form a jointly orthogonal set with $e = g \oplus h$ and $f = g \oplus i$ (such a 'minimal' orthogonal decomposition is called a Mackey decomposition).[5] Conversely, if two effects e and f are compatible in the abstract sense, they have an orthogonal decomposition (in fact a Mackey decomposition). But orthogonality in the abstract sense coincides with orthogonality in our sense above, and we have already seen that this implies compatibility also in our sense.

Using the partial order, we can finally define *sharp* elements of the effect algebra ('sharp effects' or 'projections'), as those satisfying $e \wedge e^\perp = 0$ (meaning that the greatest lower bound of e and $e^\perp$ exists and is 0). Since 0 is the minimal element of the partially ordered set (poset) $\mathcal{E}$, this in turn means that *every* lower bound of e and $e^\perp$ is 0, i.e.,

$$p(f) \leq \min[p(e), 1 - p(e)] \text{ for all } p \quad \Rightarrow \quad f = 0\,. \tag{26.50}$$

We can now show that the sharp elements of $\mathcal{E}$ form an *orthoalgebra* $\mathcal{L}$, i.e., in addition to (E1)–(E4), they satisfy also:

(E5) If $e \oplus e$ is defined, then $e = 0$

(in this case, (E4) in fact becomes redundant).

[5] Note that, once g is given, h and i are uniquely defined by $h = e \ominus g$ and $i = f \ominus g$. However, $g = [e_A \wedge f_A]$ itself is generally not independent of the choice of A, i.e., Mackey decompositions are generally not unique, so that one cannot define a partial operation $\wedge$ in this way. (Some pairs of effects may nevertheless have greatest lower bounds in the sense of the partial order.)

Proof:
Let e be sharp, and let $e \oplus e$ be defined. Then, by (26.41),

$$p(e \oplus e) = 2p(e) \quad \text{for all } p\,, \tag{26.51}$$

and thus $p(e) \leq \frac{1}{2}$ for all p. But then $1 - p(e) \geq \frac{1}{2}$ for all p, therefore

$$p(e) = \min[p(e), 1 - p(e)] \quad \text{for all } p\,. \tag{26.52}$$

Since e is sharp, $e = 0$ (by (26.50)), and (E5) follows. *QED*.

Note that the sharp elements of an arbitrary effect algebra need not form an orthoalgebra in general, so this last result depends in fact on how we have constructed $\mathcal{E}$.

Without further requirements on the experiments and the states, however, we cannot guarantee that general observables can be somehow reduced to sharp observables. (Indeed, nothing forces the orthoalgebra of sharp elements of $\mathcal{E}$ to be an interestingly rich structure — there might even be no sharp effects besides 0 and 1!)

One could ask further what conditions one might impose on $\mathcal{E}$ or $\mathcal{L}$ in order to recover classical or quantum probability. In the case of classical probability, it is obvious that $\mathcal{L}$ needs to be a Boolean algebra. In the case of quantum probability, there are some classic partial results going some way towards ensuring that the orthoalgebra of sharp effects be isomorphic to the projections on a complex Hilbert space. For instance, one might impose the following conditions on an orthoalgebra, in increasing order of strength:

(α) Unique Mackey Decomposition (UMD), i.e., compatible pairs are required to have *unique* Mackey decompositions. This ensures that the Boolean structure of sharp experiments coincides where experiments overlap, in particular allowing conjunction and disjunction to be defined globally as partial operations on the orthoalgebra (thereby turning an orthoalgebra into a so-called Boolean manifold).
(β) Orthocoherence, i.e., pairwise orthogonal sets are jointly orthogonal. This ensures that an orthoalgebra is an orthomodular poset.
(γ) Coherence, i.e., pairwise compatible sets are jointly compatible. This ensures that an orthoalgebra is a (transitive) partial Boolean algebra.

These are all properties of the orthoalgebras of sharp effects in both quantum and classical probability (indeed, we have seen this explicitly in the case of coherence).

A well-known problem, however, relates to the existence of tensor products, i.e., the possibility of composing generalized probability structures. If enough states exist (in a well-defined sense), one can construct tensor products of orthoalgebras, but forming tensor products tends to destroy orthocoherence. In this respect, the theory still needs to be investigated further.

Fortunately, these are not questions we need to address in order to discuss whether a generalization of the notion of probability such as the above is indeed necessary, or whether generalized probabilities might after all be embeddable in some suitable classical probability space. This is what we discuss in the next section, and we can do it by looking at very simple examples.

26.5 Non-Embeddability (and No-hidden-variables)

In Section 26.4, we have sketched a generalized theory of probability that is non-classical in the sense that it allows for incompatibility of observables. None of what we have said so far, however, shows that it is impossible to *embed* such a generalized probabilistic structure into some larger classical probability space, at least under some minimal assumptions.

For instance, if we require that the orthoalgebra $\mathcal{L}$ of sharp effects have the UMD property, then we can define partial Boolean operations on the sharp effects. One might now naïvely imagine *formally* extending the Boolean operations even to pairs of incompatible ones, and defining a probability measure on the thus enlarged event space that should return the original measures as marginals on each of the original Boolean algebras. Indeed, if two sharp effects e and f are not compatible, the joint probability of $e \wedge f$ might not be experimentally meaningful, but by the same token no experiment would constrain us in choosing a measure that specified also these joint probabilities (e.g. by considering e and f as independent) — or so it might seem. And if the orthoalgebra of sharp events is rich enough, one might get back all the original observables by considering unsharp realizations of sharp observables, as in the classical case of Section 26.2. Even if we considered this a purely formal construction, it would still mean that our 'generalized probability spaces' are indeed embeddable into classical probability spaces, and thus provide only a fairly trivial generalization of the formalism of classical probability.

Where this argument goes wrong, however, is in failing to realize the richness of the compatibility structure in a general orthoalgebra. The relation of compatibility is clearly reflexive and symmetric, but it is not transitive, so that observables do not fall neatly into different equivalence classes of mutually compatible ones. Put slightly differently, if compatibility is not transitive, it is possible for the same observable to be a coarse-graining of two mutually incompatible observables — which in this sense can be said to be *partially* compatible. And the interlocking structure of partially compatible observables (and the corresponding partially overlapping probability measures) can be surprisingly rich. In such cases the question of whether general probabilistic states might be induced by classical probability measures becomes non-trivial.

We shall now construct a simple master example showing explicitly that probabilistic states cannot in general be induced by classical probability measures. We shall then see how the example relates to classic impossibility theorems ruling out various kinds of 'hidden variables' theories in quantum mechanics (in particular Bell's theorem and the Kochen–Specker theorem).[6]

Imagine that we have three boxes, A, B and C. We can open any box, and find (or fail to find) a gem in it. Let e,f,g be the outcomes 'finding a gem in the box' in each of the three experiments. We further imagine that we can open any two but *not* all three boxes simultaneously. We finally imagine that we have probabilistic states specifying for any pair of boxes the probabilities for finding gems in neither, either or both of the boxes.

[6] Indeed, our example generalizes one given by Albert (1992) for the purpose of illustrating the Bell inequalities.

Now let us take a special state, such that for any ordered pair (x,y) with $x,y \in \{e,f,g\}$:

$$p(x|y) = p(\neg x|\neg y) = \alpha\,, \tag{26.53}$$

with $\alpha \in [0,1]$. For $\alpha = 1$, such a state obviously has the form

$$\begin{aligned} p(x \wedge y) &= a\,,\\ p(\neg x \wedge \neg y) &= 1-a\,,\\ p(x \wedge \neg y) &= 0\,,\\ p(\neg x \wedge y) &= 0 \end{aligned} \tag{26.54}$$

for all pairs (x,y), for some $a \in [0,1]$. And it is equally obvious that for any a, this state is (uniquely) induced by the probability measure defined by

$$p(e \wedge f \wedge g) = a\,, \quad p(\neg e \wedge \neg f \wedge \neg g) = 1-a\,, \tag{26.55}$$

and

$$\begin{aligned} p(e \wedge f \wedge \neg g) = p(e \wedge \neg f \wedge g) = p(e \wedge \neg f \wedge \neg g) = \\ p(\neg e \wedge f \wedge g) = p(\neg e \wedge f \wedge \neg g) = p(\neg e \wedge \neg f \wedge g) = 0\,. \end{aligned} \tag{26.56}$$

For the case $\alpha \neq 1$ we have instead:

Lemma:
A state satisfying (26.53) *with* $\alpha \neq 1$ *is uniquely given by*

$$p(x \wedge y) = p(\neg x \wedge \neg y) = \frac{\alpha}{2}\,, \tag{26.57}$$

and

$$p(x \wedge \neg y) = p(\neg x \wedge y) = \frac{1-\alpha}{2} \tag{26.58}$$

for any (x,y). *Thus in particular,*

$$p(x) = p(\neg x) = p(y) = p(\neg y) = \frac{1}{2}\,. \tag{26.59}$$

We then have a rather striking result:

Proposition:
Under the assumptions of the Lemma, p *is induced by a joint probability measure on the Boolean algebra (formally) generated by* $\{e,f,g\}$ *if and only if* $\alpha \geq \frac{1}{3}$.

(The proofs can be found in the Appendix of the online version of this chapter; see the note at the beginning of the chapter.) Intuitively, the case $\alpha = 1$ corresponds to perfect correlations between finding or not finding gems in any two boxes, and it is indeed obvious that this state can be extended to a probability measure in which there are perfect correlations between finding gems in all three boxes. To take an intermediate case, $\alpha = \frac{1}{2}$ is the uncorrelated case, in which any two boxes are independent, and is again obviously

extendable to the case in which all three boxes are independent. The case $\alpha = 0$ instead is the perfectly anti-correlated case, and is clearly *not* classically reproducible: if whenever there is a gem in the first box there is no gem in the second, and whenever there is no gem in the second there is one in the third, then whenever there is a gem in the first box there also is one in the third, contradicting the hypothesis. In fact, every state with $\alpha < 1$ is a convex combination of the (unique) perfectly anti-correlated state and the (special) perfectly correlated state with $p(e) = p(f) = p(g) = \frac{1}{2}$. While all states with positive correlations, the uncorrelated state, and even some with negative correlations are reproducible classically (all states with $\alpha \geq \frac{1}{3}$), if the perfectly anti-correlated component comes to dominate too strongly, the negative correlations can no longer be reproduced by a classical probability measure.

As we shall now see, a larger set of states *can*, however, be reproduced quantum mechanically (all states with $\alpha \geq \frac{1}{4}$). This shows, indeed, that already the case of quantum probabilities requires generalized probabilities that cannot be embedded in classical probability spaces. It also shows explicitly that quantum probabilities are only a subset of all possible generalized probabilities.

Take two spin-$\frac{1}{2}$ systems in the so-called singlet state,

$$\frac{1}{\sqrt{2}}(|+\rangle|-\rangle + |-\rangle|+\rangle) \,. \tag{26.60}$$

In this state, results of spin measurements in the same direction on the two electrons are perfectly anti-correlated. The singlet state (26.60) is also rotationally symmetric, so that perfect anti-correlations are obtained for pairs of parallel measurements in whatever direction. Pairs of spin measurements on different particles are always compatible, and the joint probability for spin up in direction r on the left and $-r'$ on the right is equal to $\cos^2(\vartheta/2)$, where ϑ is the angle between r and r'. Note that taking the two directions r, r'' on the left and the two directions $-r', -r$ on the right, one obtains four compatible pairs, each comprising one direction on the left and one on the right.

Now, if one is attempting to construct a joint probability measure for all four spin observables, then, given the perfect correlations for the pair of measurements in the directions $(r, -r)$, the constraint (26.53) with $\alpha = \cos^2(\vartheta/2)$ will hold of the joint probabilities for $(r, -r')$ and $(-r', r'')$ (in both cases defined on different sides) and the putative joint probabilities for (r'', r) (defined on the same side).

It is obvious that one can have three spatial directions pairwise spanning the same angle ϑ iff this angle is between 0 (when the three directions are collinear) and 120 degrees (when they are coplanar). This corresponds exactly to values of $\cos^2(\vartheta/2)$ between 1 and $\frac{1}{4}$. And it gives us a quantum model for states satisfying (26.53) with $\frac{1}{4} \leq \alpha \leq 1$, as desired.

We can now make explicit the connection with classic results in quantum mechanics about ruling out various kinds of 'hidden variables' models, in particular the Bell inequalities and the Kochen–Specker theorem.

The best-known Bell inequality is the Clauser–Holt–Shimony–Holt (CHSH) inequality (Clauser *et al.* 1969),

$$-2 \leq E(AB) - E(AB') + E(A'B) + E(A'B') \leq 2 \,, \tag{26.61}$$

where A, A', B, B' are two-valued observables with the values ± 1, each of A, A' is compatible with each of B, B', and

$$E(XY) = p(X=1, Y=1) - p(X=-1, Y=1) - p(X=1, Y=-1) + p(X=-1, Y=-1) \quad (26.62)$$

is the correlation coefficient of X and Y. As is well-known, the CHSH inequality (26.61) can be derived from the assumption of a local hidden variables model when A, A' and B, B' are interpreted, respectively, as pairs of observables pertaining to two (space-like) separated systems, e.g. spin-$\frac{1}{2}$ observables in various directions for two different particles (Bell 1971/2004). (One readily recognizes that the quantum version of our example above uses the same set-up, with a special choice of directions.)

It was Fine (1982) who first pointed out that (26.61) is also the necessary and sufficient condition for the existence of a joint probability measure for the observables A, A', B, B' when the marginals for the four compatible pairs are given. Such a joint probability measure is known as a 'non-contextual hidden variables' model of the experimental situation, since the same measure returns the correct marginals irrespective of how an observable is paired with an observable on the other side, or indeed on how an observable is assumed to be measured. Pitowsky (1989a,b) then gave a general and systematic treatment of necessary and sufficient conditions for the existence of joint probability measures in terms of such inequalities, further pointing out that such results had already been anticipated more than a century earlier by George Boole (1862).

In this sense, our discussion of the master example above must be a special case of a Bell inequality, and in fact it is a special case of (26.61). Setting $A = B'$ in (26.61) we get

$$-2 \leq E(AB) - E(AA) + E(A'B) + E(A'A) = E(AB) - 1 + E(A'B) + E(A'A) \leq 2 \quad (26.63)$$

for any three two-valued observables. Interpreting any two of them as 'finding or not finding a gem in the box', and substituting the probabilities (26.53) into (26.62), we have

$$E(XY) = 2\frac{\alpha}{2} - 2\frac{1-\alpha}{2} = \alpha - (1-\alpha) = 2\alpha - 1 \quad (26.64)$$

for any distinct X, Y, and (26.63) becomes

$$-2 \leq 6\alpha - 3 - 1 \leq 2\,, \quad (26.65)$$

that is

$$2 \leq 6\alpha \leq 6\,. \quad (26.66)$$

Thus $\alpha \in [\frac{1}{3}, 1]$, as above.

The Kochen–Specker theorem instead takes finite sets of projection-valued observables in a Hilbert space of dimension at least 3, that may pairwise share a projection (partially compatible observables), and considers the question of whether values 1 and 0 may be assigned to the projections in such a way that exactly one projection from each observable is assigned the value 1. (The theorem was first announced in Specker (1960), and its proof was published in Kochen and Specker (1967/1975).)

One already knows that making such assignments to *all* projection-valued observables in a Hilbert space of dimension at least 3 must lead to a contradiction. Indeed, such assignments are simply trivial probability measures over the projections in Hilbert space, and by Gleason's theorem (which we discussed in Section 26.3) the most general such probability measures are the quantum mechanical states, which in fact always assign non-trivial probabilities to *some* observables. By the compactness theorem of first-order logic, one then knows also that there must be a finite set of observables for which such an assignment leads to a contradiction. But Kochen and Specker offer a constructive proof that such a finite set of observables exists (the original proof involved 117 one-dimensional projections in 3 dimensions).

The case $\alpha = 0$ in our example can now be seen as a 'two-dimensional' analogue of the Kochen–Specker theorem. Indeed, it can be seen as comprised of three interlocking pairs of projections, such that exactly one element in each pair is assigned the value 1, and the other one the value 0.

The analogy goes both ways: any Kochen–Specker construction (of finite sets of orthonormal bases that cannot be assigned values 1 and 0 in such a way that exactly one vector in each basis is assigned 1) is equivalent to the non-existence of trivial probability measures satisfying suitable constraints (in three dimensions, these are $p(x \vee y|z) = 0$ and $p(\neg x|\neg y \wedge \neg z) = 0$ for all orthonormal triples). But the existence of trivial probability measures satisfying such constraints is in fact equivalent to the existence of non-trivial probability measures satisfying the same constraints. Thus, indeed, every Kochen–Specker theorem can be translated into the violation of some Bell–Pitowsky inequality.[7]

26.6 Is Probability Empirical (and Quantum)?

The title of this section recalls (tongue-in-cheek) the title of the classic paper by Putnam (1968) in which he notoriously argued that quantum mechanics requires a fundamental revision of logic. Empirical considerations alone presumably cannot decide the question of whether logic is an empirical or an a priori discipline (as forcefully pointed out in another classic paper by Dummett (1976)). But if one is already sympathetic to the idea that logic is an empirical discipline, then it does make sense to ask what kind of empirical evidence might suggest adopting this or that logic, and in particular whether the evidence we have for quantum mechanics suggests adopting a non-classical one (e.g. one based on Kochen and Specker's partial Boolean algebras). Essentially, the question boils down to whether quantum logic should be seen as a derivative construct that is definable in terms of and alongside classical logic, or whether classical logic should be seen as an instance of quantum logic restricted to certain special 'well-behaved' cases.[8]

[7] For a powerful and systematic treatment of these issues, see Abramsky and Brandenburger (2011, esp. Sect. 7).

[8] For a recent review, emphasizing that the answer might depend rather sensitively on one's chosen interpretation of quantum mechanics, see Bacciagaluppi (2009).

A somewhat similar question might be asked with regard to probability. We have seen in Sections 26.3–26.4 that quantum mechanics suggests introducing a notion of probabilistic state generalizing that of a probability measure, and the non-embeddability result we have derived in Section 26.5 states that in general the joint probability distributions defined by a state (in particular a quantum-mechanical state) for certain pairs of observables cannot be recovered as marginals of a single classical probability measure. In this final section, we shall discuss whether these results should compel us to see classical probabilities as a special case of generalized probabilities (for the case in which all observables are compatible), or whether generalized probability theory could after all be derivative of classical probability.

There is a sense in which the latter question can indeed be trivially answered in the affirmative, by taking seriously the idea that a general probabilistic state should be seen as a family of classical probability measures, but denying that they in fact overlap. For instance, in our master example above (say, with $\alpha = 0$), this means simply that instead of describing the relevant probabilistic structure using a single state that assigns the probabilities

$$p(e) = p(f) = p(g) = \frac{1}{2} \tag{26.67}$$

to the outcomes e, f and g (and the appropriate joint probabilities to pairs), we describe it using *three different* classical probability measures, which are to be applied respectively to the experiments in which we measure e and f together, or f and g together, or g and e together, and that assign, respectively, the probabilities

$$p_{ef}(e) = p_{ef}(f) = \frac{1}{2}, \quad p_{fg}(f) = p_{fg}(g) = \frac{1}{2}, \quad p_{ge}(g) = p_{ge}(e) = \frac{1}{2}, \tag{26.68}$$

to the single outcomes (and the appropriate joint probabilities to the three pairs). We see now that these three probability measures can be derived from a single classical probability measure if we assign probabilities also to performing each of the experiments ef, fg and ge. This is just the 'naïve' argument we rehearsed at the beginning of Section 26.5, but which is now no longer blocked, because we resist identifying the two events e_{ef} and e_{ge} as one (and similarly for f and for g). In this (formal) sense, a 'contextual hidden variables theory' is always possible. (Note, however, that if we then imagine performing the three joint measurements ef, fg and ge in sequence, then the measured value of at least one observable, say e, must be different in the two measurements containing it, in this case ef and ge. Thus we have some mysterious form of 'disturbance through measurement', much as in our simple discussion of spin in Section 26.1.)

What is crucial here is that instead of insisting that experimental outcomes belonging to the same effect be identified as the same event, we rather insist that experimental outcomes belonging to different experiments are *different events*. This suggestion should not be too hastily dismissed.[9] Identifying experimental outcomes that are equiprobable in all states might after all be thought of only as a convenient book-keeping device of no fundamental importance. (Even in quantum mechanics, as we pointed out in Section 26.3, the same effect can correspond to different physical transformations of the state in different experiments, so one could very well argue that these should be considered different events.)

[9] In this connection, see also Dzhafarov and Kujala (2014a,b).

In quantum mechanics, these questions are played out in the context of the debate on hidden variables theories (see e.g. Shimony 1984). If effects (or at least projections) correspond directly to physical properties of a quantum system, which are then measured in various ways, then projections in common to different observables (resolutions of the identity) should, indeed, be identified. If instead the properties of the system are some 'hidden variables', which in the context of specific experimental arrangements lead to certain experimental outcomes (perhaps with certain probabilities), then projections no longer represent intrinsic properties of the system in general, but only aspects of how systems can be probed in the context of specific experimental situations.

These brief remarks suggest that the question of whether different experimental outcomes ought to be identified should not be decided abstractly, but rather in relation to specific theoretical commitments. We shall not attempt a general discussion of this point, nor even an exhaustive one of hidden variables theories in quantum mechanics. What we shall do instead is illustrate the point in some concrete implementations of our master example (mainly for $\alpha = 0$), which will enable us to see the possibility of underlying mechanisms providing us with a rationale for deciding when different experimental outcomes should be treated as different events.

The original setting of our case $\alpha = 0$ (but without the probabilistic structure) is the tale of the Sage of Nineveh from Specker (1960) (my translation):[10]

> At the Assyrian school for prophets in Arba'ilu, there taught, in the age of king Asarhaddon, a sage from Nineveh. He was an outstanding representative of his discipline (solar and lunar eclipses), who, except for the heavenly bodies, had thoughts almost only for his daughter. His teaching success was modest; the discipline was seen as dry, and furthermore required previous mathematical knowledge that was rarely available. If in his teaching he thus failed to get the interest he would have wanted from the students, he received it overabundantly in a different field: no sooner had his daughter reached marriageable age, than he was flooded with requests for her hand from students and young graduates. And even though he did not imagine wishing to keep her with him for ever, yet she was still far too young, and the suitors in no way worthy of her. And so that each himself should be assured of his unworthiness, he promised her hand to the one who could perform a set prophecy task. The suitor was led in front of a table on which stood three boxes in a row, and urged to say which boxes contained a gem and which were empty. Yet, as many as would try it, it appeared impossible to perform the task. After his prophecy, each suitor was indeed urged by the father to open two boxes that he had named as both empty or as both not empty: it always proved to be that one contained a gem and the other did not, and actually the gem lay now in the first, now in the second of the opened boxes. But how should it be possible, out of three boxes, to name no two as empty or as not empty? Thus indeed the daughter would have remained unmarried until her father's death, had she not upon the prophecy of a prophet's son swiftly opened two boxes herself, namely one named as full and one named as empty — which they yet truly turned out to be. Upon the father's weak protest that he wanted to have two different boxes opened, she tried to open also the third box, which however proved to be impossible, upon which the father, grumbling, let the unfalsified prophecy count as successful.

[10] Specker was a great story-teller, and I personally heard him tell this particular story I think in the spring of 1985.

The question we wish to address is whether opening box A in the context of also opening box B is the same event as opening box A in the context of also opening box C. Given our usual intuitions, i.e., background theoretical assumptions, it would seem that we do have to identify the two events. But we can also imagine the situation as follows.

What the father wants to establish is whether any of the suitors are better prophets than himself (only then would he willingly surrender his daughter's hand in marriage). Whenever a suitor is set the task, the father predicts which two boxes will be opened, and places exactly one gem at random in one of the two boxes. (Note that, in this form, the example bears some analogy to Newcomb's paradox!) If we now assume that the father possesses a genuine gift for clairvoyant prophecy, the action of opening boxes A and B, or of opening A and C, has a retrocausal effect on whether the father has placed the gem in either A or B, or has placed it in either A or C.

Now we have an explanation of why to each of the three experimental situations corresponds a *different* classical probability measure, and there appears to be no longer a motivation for describing the situation using a single probabilistic state irrespective of the experimental situation. (Note that should probabilities be defined also for which two of the boxes will be opened, one could again introduce a single classical measure from which the three probability measures arise through conditionalization.)

We can imagine a different mechanism (and arrive at the same conclusion) by considering another classic illustration of our example, the so-called 'firefly box'. Imagine that the three boxes are in fact chambers at the corners of a single box in the shape of an equilateral triangle, all three chambers being accessible from the centre. Assume that we are in darkness, and hold up a lantern to any one of the sides of the triangle. And assume that what we observe is always that (at random) one chamber on the illuminated side faintly starts to glow. Probabilistically, this is exactly the same example as above. But we can now imagine a different explanation for this phenomenon, as follows.

At the centre of the box sits a firefly, which is attracted to the light of our lantern, and thus enters at random one of the two chambers on the side from which we are approaching. And mistaking our lantern for a potential mating partner, the firefly starts to glow!

We have again a mechanism explaining the statistics of our experiments, and we can give the same classical probabilistic model of the situation as before, i.e., we have three different experiments, each of which is described by a different classical probability measure. And if we so wish, we can again introduce probabilities for our approaching from any particular side.

Quantum mechanics provides us with non-classical, non-contextual probabilistic models of various phenomena, and several impossibility theorems show that there fail to be any classical non-contextual probabilistic models reproducing the quantum mechanical statistics (so-called non-contextual hidden variables theories). In these examples instead we see the analogues of various strategies used in quantum mechanics to introduce *classical but contextual* probabilistic models (so-called contextual hidden variables theories).

Indeed, only few retrocausal models of quantum mechanics have been developed in detail, but retrocausality has long been recognized as a possible strategy to deal with the puzzles of quantum mechanics, in particular in the face of the Bell inequalities. The firefly model instead more closely resembles a theory like de Broglie and Bohm's pilot-wave theory, in which experimental outcomes depend on both the initial configuration of the system (e.g.

the position of an electron) and the details of the experimental arrangement. This last point can probably best be seen in another slight variant of the example.

Imagine that instead of the firefly we have a small metal ball in the centre of the box, and that each experiment consists of tilting the box towards one of the sides, say AB. The ball rolls towards the side AB and bounces off a metal pin into either chamber A or chamber B, depending on its exact initial location to the left or the right of the symmetry axis perpendicular to the side AB. It is now clear that the *same* initial position of the ball might lead it to fall or not to fall into, say, chamber A, depending on whether the whole box is tilted towards the side AB or the side CA (namely if the ball is on the left of the symmetry axis through AB *as well as* to the left of the symmetry axis through CA). Thus, depending on which way the box is tilted, the ball ending up in A corresponds to a *different* random variable on the probability space of initial positions of the ball. If the initial position of the ball is uniformly distributed in a symmetric neighbourhood of the centre of the triangle, the equal probabilities of the non-classical state are reproduced. But if the initial position is not in such an 'equilibrium' distribution, deviations from the probabilities in the Lemma can occur — so that if one allows also such 'disequilibrium' hidden states, experimental outcomes are in fact no longer equiprobable in all states.

Our examples can be easily generalized to include e.g. the original Kochen–Specker example — for which we need a spherical firefly box with 117 sub-chambers, only three of which are made accessible to the firefly every time we approach the box (depending on how exactly we approach it). Or indeed to include cases with $\alpha > 0$. For the latter, we need a cubical firefly box, which we approach from any of the six faces (counting opposite faces as equivalent). On each face, the four corners correspond to, say, $e \wedge f$ and $\neg e \wedge \neg f$ across one diagonal, and $e \wedge \neg f$ and $\neg e \wedge f$ across the other, and similarly with f and g, or g and e, on the other faces. The classical cases can be obtained if the firefly just sits somewhere in the box (maybe preferentially along one spatial diagonal — where food might be provided), and starts to glow when it sees the light from our lantern. We then observe the projections of the firefly's position on the face from which we approach. The non-classical cases can be obtained if the firefly moves towards the side from which we are approaching, and through various obstacles is channelled preferentially (although not always) along the planar diagonal corresponding to the opposite outcomes for that face (say $e \wedge \neg f$ and $\neg e \wedge f$). We can thus construct classical but contextual models that violate (our special case of) the Bell inequalities, reproducing the quantum violations, or even the non-quantum violations (reducing to the equilateral triangle in the limit).

In conclusion, while the results of Section 26.5 show that a generalized probabilistic model as introduced in Section 26.4 cannot always be embedded in a single classical probability space, the examples in this section indicate that it can always be reproduced using a family of classical probability measures indexed by different experimental contexts, *if* indeed we have a reason to resist the temptation to identify experimental outcomes across different experiments. Identifying experimental outcomes that are equiprobable in all states may be completely natural once a theoretical setting is given; but whether two events are to be judged the same is not a formal question, nor can it be decided purely on operational grounds. Instead it depends on the choice of theoretical setting. In the specific case of quantum probabilities, this question is closely related to the notorious question of the interpretation of quantum mechanics.

Acknowledgments

I would like to thank the editors for their invitation to contribute to this volume and the opportunity to write on this topic, but very especially for their saintly patience and for very helpful feedback on a previous version of this chapter. I am further grateful to Alex Wilce for some extremely useful discussions and hard-to-find references, to Jennifer Bailey for some stylistic advice, and to the audience of the Philosophy of Physics seminar at the University of Aberdeen, who heard preliminary versions of this material.

References

Abramsky, S. and Brandenburger, A. (2011) The Sheaf-theoretic Structure of Non-locality and Contextuality. *New Journal of Physics*. 13. 11. 113036. [Online] Available from: http://arxiv.org/abs/1102.0264. [Accessed: 24 Aug 2015].

Albert, D. (1992) *Quantum Mechanics and Experience*. Cambridge, MA: Harvard University Press.

Bacciagaluppi, G. (2009) Is Logic Empirical? In Gabbay, D., Lehmann, D., and Engesser, K. (eds.) *Handbook of Quantum Logic*. pp. 49–78. Amsterdam: Elsevier.

Bacciagaluppi, G. (2013) Measurement and Classical Regime in Quantum Mechanics. In Batterman, R. (ed.) *Oxford Handbook of Philosophy of Physics*. pp. 416–59. Oxford: Oxford University Press.

Bell, J. S. (1971/2004) Introduction to the Hidden-Variable Question. In *Proceedings of the International School of Physics 'Enrico Fermi', Course IL, Foundations of Quantum Mechanics*. pp. 171–81. New York, NY: Academic. (Reprinted in Bell, J. S. *Speakable and Unspeakable in Quantum Mechanics*. pp. 29–39. Cambridge: Cambridge University Press.)

Beltrametti, E. G. and Cassinelli, G. (1981) *The Logic of Quantum Mechanics*. Reading, MA: Addison-Wesley.

Birkhoff, G. and von Neumann, J. (1936/1975) The Logic of Quantum Mechanics. *Annals of Mathematics*. 37. pp. 823–43. (Reprinted in Hooker, C. A. *The Logico-Algebraic Approach to Quantum Mechanics*. Vol. 1. pp. 1–26. Dordrecht: Reidel.)

Boole, G. (1862) On the Theory of Probabilities. *Philosophical Transactions of the Royal Society of London*. 152. pp. 225–52.

Busch, P. (2003) Quantum States and Generalized Observables: A Simple Proof of Gleason's Theorem. *Physical Review Letters*. 91 12. 120403. [Online] Available from: http://arxiv.org/abs/quant-ph/9909073. [Accessed 24 Aug 2015].

Busch, P., Grabowski, M., and Lahti, P. J. (1995) *Operational Quantum Physics*. Berlin: Springer. (2nd, corrected printing 1997.)

Busch, P., Lahti, P. J., and Mittelstaedt, P. (1991) *The Quantum Theory of Measurement*. Berlin: Springer (2nd edition 1996.)

Cattaneo, G., Marsico, T., Nisticò, G., and Bacciagaluppi, G. (1997) A Concrete Procedure for Obtaining Sharp Reconstructions of Unsharp Observables in Finite-Dimensional Quantum Mechanics. *Foundations of Physics*. 27. pp. 1323–43.

Clauser, J. F., Horne, M. A., Shimony, A., and Holt, R. A. (1969) Proposed Experiment to Test Local Hidden-Variable Theories. *Physical Review Letters*. 23. 15. pp. 880–4.

de Barros, J. A. and Suppes, P. (2010) Probabilistic Inequalities and Upper Probabilities in Quantum Mechanical Entanglement. *Manuscrito.* 33. 1. pp. 55–71. [Online] Available from: http://arxiv.org/abs/1010.3064 [Accessed 24 Aug 2015].

Dummett, M. (1976) Is Logic Empirical? In Lewis, H. D. (ed.) *Contemporary British Philosophy*. 4th series pp. 45–68. London: Allen and Unwin.

Dzhafarov, E. and Kujala, J. V. (2014a) Probabilistic Contextuality in EPR/Bohm-type Systems with Signaling Allowed. [Online] Available from: http://arxiv.org/abs/1406.0243. [Accessed 24 Aug 2015].

Dzhafarov, E. and Kujala, J. V. (2014b) Generalizing Bell-type and Leggett–Garg-type Inequalities to Systems with Signaling. [Online] Available from: http://arxiv.org/abs/1407.2886. [Accessed 24 Aug 2015]

Fine, A. (1982) Hidden Variables, Joint Probability and Bell Inequalities. *Physical Review Letters*. 48. pp. 291–5.

Foulis, D. J. and Randall, C. H. (1972) Operational Statistics. I. Basic Concepts. *Journal of Mathematical Physics*. 13. pp. 1667–75.

Foulis, D. J. and Randall, C. H. (1974) Empirical Logic and Quantum Mechanics. *Synthese*. 29. pp. 81–111.

Ghirardi, G. C. (1997) *Un'occhiata alle carte di Dio*. Milan: Il Saggiatore. (Translated by G. Malsbary as *Sneaking a Look at God's Cards.* Princeton: Princeton University Press. 2005.)

Gleason, A. M. (1957/1975) Measures on the Closed Subspaces of a Hilbert Space. *Journal of Mathematics and Mechanics.* 6. pp. 885–93. (Reprinted in Hooker, C. A. *The Logico-Algebraic Approach to Quantum Mechanics*. pp. 123–33. Dordrecht: Reidel.)

Kochen, S. and Specker, E. P. (1965/1975a) Logical Structures Arising in Quantum Theory. In Addison, L., Henkin, L., and Tarski, A. (eds.) *The Theory of Models*. pp. 177–89. Amsterdam: North-Holland. (Reprinted in Hooker, C. A. *The Logico-Algebraic Approach to Quantum Mechanics*. pp. 263–76. Dordrecht: Reidel.)

Kochen, S. and Specker, E. P. (1965/1975b) The Calculus of Partial Propositional Functions. In Bar-Hillel, Y. (ed.) *Logic, Methodology, and Philosophy of Science*. pp. 45–57. Amsterdam: North-Holland. (Reprinted in Hooker, C. A. *The Logico-Algebraic Approach to Quantum Mechanics*. pp. 277–92. Dordrecht: Reidel.)

Kochen, S. and Specker, E. P. (1967/1975) The Problem of Hidden Variables in Quantum Mechanics. *Journal of Mathematics and Mechanics.* 17. pp. 59–88. (Reprinted in Hooker, C. A. *The Logico-Algebraic Approach to Quantum Mechanics*. pp. 293–328. Dordrecht: Reidel.)

Ludwig, G. (1954) *Die Grundlagen der Quantenmechanik*. Berlin: Springer (Translated by C. A. Hein as *Foundations of Quantum Mechanics*. Berlin: Springer. 1983.)

Ludwig, G. (1985) *An Axiomatic Basis of Quantum Mechanics. 1. Derivation of Hilbert Space*. Berlin: Springer.

Mackey, G. W. (1957) Quantum Mechanics and Hilbert Space. *American Mathematical Monthly*. 64. 2. pp. 45–57.

Mackey, G. W. (1963) *The Mathematical Foundations of Quantum Mechanics*. New York: W. A. Benjamin.

Oas, G., de Barros, J. A., and Carvalhaes, C. (2014) Exploring Non-signalling Polytopes with Negative Probability. [Online] Available from: http://arxiv.org/abs/1404.3831. [Accessed 24 Aug 2015.]

Pitowsky, I. (1989a) From George Boole to John Bell: The Origins of Bell's Inequality. In Kafatos, M. (ed.) *Bell's Theorem, Quantum Theory and the Conceptions of the Universe* pp. 37–49. Dordrecht: Kluwer.

Pitowsky, I. (1989b) *Quantum Probability, Quantum Logic*. Lecture Notes in Physics. Vol. 321. Berlin: Springer.

Putnam, H. (1968) Is Logic Empirical? In Cohen, R. and Wartofsky, M. (eds.) *Boston Studies in the Philosophy of Science*. Vol. 5. pp. 216–41. Dordrecht: Reidel.

Randall, C. H. and Foulis, D. J. (1970) An Approach to Empirical Logic. *American Mathematical Monthly*. 77. pp. 363–74.

Randall, C. H. and Foulis, D. J. (1973) Operational Statistics. II. Manuals of Operations and their Logics. *Journal of Mathematical Physics*. 14. pp. 1472–80.

Shimony, A. (1984) Contextual Hidden Variables Theories and Bell's Inequalities. *The British Journal for the Philosophy of Science*. 35. 1. pp. 25–45.

Specker, E. P. (1960) Die Logik nicht gleichzeitig entscheidbarer Aussagen. *Dialectica*. 14. pp. 239–46. (Translated by A. Stairs as The Logic of Propositions which are not Simultaneously Decidable. In Hooker, C. A. (1975) *The Logico-Algebraic Approach to Quantum Mechanics*. pp. 135–40.)

Varadarajan, V. S. (1968) *The Geometry of Quantum Theory*. Vols. I–II. New York, NY: Van Nostrand.

Wallace, D. (2008) Philosophy of Quantum Mechanics. In Rickles, D. (ed.) *The Ashgate Companion to Contemporary Philosophy of Physics*. pp. 16–98. Aldershot: Ashgate.

Wilce, A. (2000) Test Spaces and Orthoalgebras. In Coecke, B., Moore, D., and Wilce, A. (eds.) *Current Research in Operational Quantum Logic*. pp. 81–114. Dordrecht: Kluwer.

CHAPTER 27

PROBABILITIES IN STATISTICAL MECHANICS

WAYNE C. MYRVOLD

27.1 Introduction

Probabilities first entered physics in a systematic way in connection with the kinetic theory of gases, according to which a gas consists of a large number of molecules moving about in a haphazard, effectively random way.[1] This theory was developed, at the hands of Maxwell, Boltzmann, and Gibbs, into the science that we (following Gibbs) now call *statistical mechanics*, a theory whose scope has been extended well beyond treatment of gases. Though statistical mechanics has grown into a well-established branch of physics with a substantial array of agreed-upon techniques of calculation, with impressive empirical success, there is little agreement on the ultimate rationale for its methods. For this reason, there has arisen a substantial philosophical literature on conceptual issues associated with statistical mechanics. Much of the philosophical discussion deals, in one way or another, with the role of probability in statistical mechanics.

This chapter will review selected aspects of the terrain of discussions about probabilities in statistical mechanics (with no pretensions to exhaustiveness, though the major issues will be touched upon), and will argue for a number of claims. None of the claims to be defended is entirely original, but each deserves emphasis. The first, and least controversial, is that probabilistic notions are needed to make sense of statistical mechanics. The reason for this—which was, in fact the reason that convinced Maxwell, Gibbs, and Boltzmann that probabilities would be needed—is that the second law of thermodynamics, which in its original formulation says that certain processes are impossible, must, on the kinetic theory, be replaced by a weaker formulation according to which what the original version deems impossible is merely improbable. The second is that we ought not to take the standard measures invoked in equilibrium statistical mechanics as giving, in *any* sense, the correct probabilities about microstates of the system. We can settle for a much weaker claim: that

[1] See Brush (1976a,b) for a survey of the early history.

the probabilities for outcomes of experiments yielded by the standard distributions are effectively the same as those for those yielded by any distribution that we *should* take as representing probabilities over microstates. Lastly (and most controversially): in asking about the status of probabilities in statistical mechanics, the familiar dichotomy between epistemic probabilities (credences, or degrees of belief) and ontic (physical) probabilities is insufficient; the concept of probability that is best suited to the needs of statistical mechanics is one that combines epistemic and physical considerations.

Outline of the chapter. I will set the stage by briefly reviewing the backdrop, in probability theory, against which the founders of statistical mechanics were working. We will then see how probabilities were introduced into statistical mechanics, and review the considerations that led Maxwell, Gibbs, and Boltzmann to conclude that probabilities were needed in statistical mechanics. Since probabilities play somewhat different roles in the two approaches to statistical mechanics that have their roots in the work of Boltzmann and Gibbs, respectively, I will briefly present these approaches. I will then discuss some approaches to justifying the choice of probability measures. Lastly, I will discuss some puzzling aspects of the use of the standard equilibrium measures, and argue that these puzzles can be resolved either by invoking quantum probabilities, or by construing the probabilities in statistical mechanics as *almost objective* probabilities.

27.2 Meanings of "Probability"

As Hacking (1975) has amply demonstrated, from the early days of probability theory, there were two distinct concepts that went by the name of "probability." One is an epistemic concept, having to do with degrees of belief. The other, which Hacking calls the *aleatory* conception, attributes probabilities to events in the world, such as the toss of a coin, which they are thought to possess independently of our knowledge or belief.[2] These need not be regarded as rivals; they are two potentially useful senses in which the word "probability" is used, which we must be careful not to conflate. Owing, perhaps, to concerns about the compatibility of objective chance with the presumed determinism of the laws of nature, we find, in the latter half of the 19th century, a frequency conception replacing single-case chance as the favored construal of objective probability. So thoroughly did the notion of single-case chance drop out of discussions that subjectivists in the first half of the 20th century, such as de Finetti and Savage, when arguing against objective notions of probability, omitted it from their lists of notions to be considered and rejected, and Popper (1957, 1959) took himself to be introducing an entirely new idea when he introduced the notion of single-case objective probabilities, which he called *propensities*.

A central question in understanding the use of probabilities in statistical mechanics is the status of these probabilities. Which notion is in play? Are they epistemic, having to do with our state of knowledge or belief about the system, or are they ontic, properties of the

[2] For one clear statement that there are two distinct senses of "probability," and a characterization of the objective notion as single-case chance, see Poisson (1837, p. 31), quoted in Myrvold (2012a, pp. 74–5).

physical systems themselves? If ontic, should they be thought of in frequentist terms, or in terms of single-case chances?

Textbook introductions of probabilities in statistical mechanics typically begin with the observation that, though the systems considered have very many degrees of freedom, our knowledge of the state of a system is limited to the measured values of a small number of thermodynamic variables. This suggests an epistemic reading of the probabilities. And, indeed, there is a long history of construing statistical mechanical probabilities as purely epistemic.[3] This fits uncomfortably, however, with the idea that the theory to be developed belongs strictly to physics, and that objective laws of thermodynamics are to be recovered on its basis. These considerations suggest an ontic reading. But, in the context of classical physics, with its deterministic laws of motion, a reading of the probabilities as objective chances seems out of the question. This seems to leave some sort of frequentism as the only option for an ontic reading of probabilities in statistical mechanics.

Is frequentism a viable option? It is certainly true that there is a close connection between frequency and chance. If I draw at random from an urn, with each ball in the urn having an equal chance of being drawn, then the chance that the drawn ball is black is equal to the relative frequency of black balls in the urn. In an infinite sequence of independent trials (such as, say, repeated rolls of a die), in each of which a certain outcome has a chance p of occurring in each trial, it can be proven (this is the Strong Law of Large Numbers) that the relative frequency of that outcome will, with chance equal to one, converge to p. Moreover, if we have available to us a long sequence of independent trials of events with equal chances, relative frequency data can be used as evidence about the values of these chances.

These considerations do not, however, enable us to *define* chance in terms of frequencies, as even to state any of them requires a notion of chance distinct from that of relative frequencies. Though the point remains somewhat controversial, there are good reasons to think that frequentism is an inadequate foundation for objective probability.[4] In light of considerations such as these, absence of consensus about the status of statistical mechanical probabilities is unsurprising. Neither an epistemic nor an ontic reading seems to be adequate for the job, at least as far as classical statistical mechanics is concerned.

One position that has been adopted is that classical statistical mechanics, rather than being an autonomous science, must borrow its probabilities from quantum mechanics. Though the determinism of classical physics undermined the notion of objective chance, quantum mechanics revived it, as quantum mechanics is often regarded as a fundamentally chancy theory. Can we think of the probabilities used in classical statistical mechanics as quantum mechanical in origin?

This possibility will be treated in §27.8.2.3. But, I will argue, such a move is not necessary. Though neither a purely epistemic nor a purely ontic reading of probabilities of statistical mechanics is available in the context of classical physics, the epistemic/ontic dichotomy is not exhaustive. As will be argued in §27.8.2.2, and in §27.9, there is a notion of probability that combines epistemic and physical considerations, that seems to be well suited for the role required of it by statistical mechanics.

[3] See Uffink (2011).

[4] For further discussion, see Jeffrey (1992), Hájek (1997, 2009).

27.3 The Introduction of Probability into Statistical Mechanics

It is useful to distinguish between two sorts of use of the probability calculus. One sort, which we may call *quasi-deterministic*, uses, implicitly or explicitly, a version of the Law of Large Numbers to replace, when dealing with a large collection of things, some quantity by its expectation value. For example, in a long enough sequence of tosses of a fair coin, we may take the fraction of tosses that are heads to be one-half, as this will, with high probability, be a good approximation. The hallmark of this use is effective certainty from uncertainty; a large number of individually unpredictable events combine to yield a result that is almost certain. The second sort of use is found in cases in which the deviations of a quantity from its expectation value are not negligible.

Von Plato (1994, p. 72) credits Krönig (1856) with the first use of probability in the context of the kinetic theory of gases. Krönig's use of probability is of the quasi-deterministic sort, to conclude that out of the irregular motion of molecules would arise regularity Krönig 1856, p. 316. We also find a quasi-deterministic use, in passing, in Clausius (1857, pp. 371–72; 1966, p. 126).

Unlike his predecessors, Maxwell did not replace a gas of molecules moving at different speeds with one in which all have the same speed, but, rather, investigated the distribution of velocities one should expect to find among the molecules of a gas. Much of his work in kinetic theory is concerned with showing that the distribution of velocities in a gas will be what is now called the Maxwell-Boltzmann distribution. In 1867 he attempted to show that molecular collisions would lead to the Maxwell-Boltzmann distribution of velocities. The argument relies on the assumption (invoked without comment) that the Ehrenfests would later call the *Stoßzahlansatz* (Ehrenfest and Ehrenfest, 1912). This is the assumption that, for pairs of molecules about to collide, one can assume probabilistic independence of the incoming velocities, and moreover, treat the two molecules as if their velocities are randomly sampled from the distribution of velocities in the gas as a whole. Maxwell shows that the Maxwell-Boltzmann distribution is stationary under collisions, and concludes that this distribution is what collisions will lead to.

Maxwell's use of probabilistic reasoning is of the quasi-deterministic sort. But it was the Maxwell-Boltzmann distribution, which makes it clear that there will be variations in speeds among the molecules of the gas, that led him eventually to conclude that the second law of thermodynamics would hold, at best, with high probability for macroscopic systems.

Boltzmann, in 1868 and, more significantly, in 1872, sought to provide a derivation more satisfactory than Maxwell's. In 1872 he argued that molecular collisions would lead to a decrease in a quantity that he called H, a result known as Boltzmann's H-theorem. Though the proof requires the *Stoßzahlansatz*, it is not highlighted by Boltzmann as a special assumption.

So far, these are all quasi-deterministic, or order-from-disorder applications of probability. However, if thermodynamic relations are relations between expectation values of quantities defined as averages of molecular properties, then we should expect to find deviations from these relations, though the probability of significant deviations will diminish as the number of molecules increases. Of particular significance is the recognition

that the second law of thermodynamics, as originally conceived, cannot, on the kinetic theory, be strictly correct; at best we can expect it to hold, with high probability, to a high degree of approximation, for systems of many degrees of freedom.

Recognition of limitations on the validity of the second law of thermodynamics appears in Maxwell's correspondence from about 1867. The key consideration is the issue of reversibility. On the assumption that intermolecular forces depend on only their relative positions, the dynamical laws governing molecular motions will be symmetric under time reversal. Therefore, thermodynamic irreversibility cannot be a consequence of dynamical considerations alone. Maxwell's view is that processes that, from the point of view of thermodynamics, are regarded as irreversible, are ones whose temporal inverses are not impossible, but merely improbable. In a letter to the editor of the *Saturday Review*, dated April 13, 1868, Maxwell draws an analogy between the mixing of fluids and balls shaken in a box.

> As a simple instance of an irreversible operation which (I think) depends on the same principle, suppose so many black balls put at the bottom of a box and so many white above them. Then let them be jumbled together. If there is no physical difference between the white and black balls, it is exceedingly improbable that any amount of shaking will bring all the black balls to the bottom and all the white to the top again, so that the operation of mixing is irreversible unless either the black balls are heavier than the white or a person who knows white from black picks them and sorts them.
>
> Thus if you put a drop of water into a vessel of water no chemist can take out that identical drop again, though he could take out a drop of any other liquid. (in Garber et al. 1995, pp. 192–3)

We find similar considerations in Gibbs several years later.

> [W]hen such gases have been mixed, there is no more impossibility of the separation of the two kinds of molecules in virtue of their ordinary motions in the gaseous mass without any external influence, than there is of the separation of a homogeneous gas into the same two parts into which it has once been divided, after these have once been mixed. In other words, the impossibility of an uncompensated decrease of entropy seems to be reduced to improbability. (Gibbs 1875, p. 229; 1906/1961, p. 167)

It is one thing to acknowledge that violations of the second law will sometimes occur, albeit with low probability. Maxwell went further, asserting that, on the small scale, minute violations of the second law will continually occur; it is only large-scale, observable violations that are improbable.

> [T]he second law of thermodynamics is continually being violated, and that to a considerable extent, in any sufficiently small group of molecules belonging to a real body. As the number of molecules in the group is increased, the deviations from the mean of the whole become smaller and less frequent; and when the number is increased till the group includes a sensible portion of the body, the probability of a measurable variation from the mean occurring in a finite number of years becomes so small that it may be regarded as practically an impossibility.
>
> This calculation belongs of course to molecular theory and not to pure thermodynamics, but it shows that we have reason for believing the truth of the second law to be of the nature of a strong probability, which, though it falls short of certainty by less than any assignable quantity, is not an absolute certainty. (Maxwell 1878b, p. 280; Niven 1890, pp. 670–1)

William Thomson (1874) provided calculations of the probability of a variety of fluctuations away from the equilibrium state of mixed gases.

In Boltzmann's work, the quasi-deterministic use of probability in his derivation of the H-theorem, together with the tacit nature of his employment of the key probabilistic assumption, the *Stoßzahlansatz*, fostered the impression that the H-theorem followed from molecular dynamics alone. As we have already noted, Maxwell and the British physicists working on kinetic theory were by this time (1872) keenly aware that there could be no derivation of an irreversible relaxation to equilibrium on the basis of reversible dynamics; in their view, probabilistic assumptions would be needed, and the conclusion to be derived would be that evolution away from macroscopic equilibrium, rather than towards it, is at best improbable, not impossible. There are no hints of reservations of this sort in Boltzmann's work of 1872. It was Loschmidt who, in 1876, drew Boltzmann's attention to reversibility considerations. In his response to Loschmidt, Boltzmann (1877a) acknowledged that there could be no purely dynamical proof of the increase of entropy.[5]

Thus, in the decade from 1867 to 1877, the major figures involved in the development of statistical mechanics concluded, on the basis of the reversibility argument, that the second law of thermodynamics, as originally conceived, could not be strictly true, and that it must be replaced by a probabilistic version, in which what is deemed impossible in the original version becomes improbable.

27.4 Revising Thermodynamics

On the kinetic theory, heat is not a substance, and it makes no sense to talk of the heat content of a body. Instead, we distinguish between two modes in which energy may be transferred from one body to another; as heat, or as work done on (or by) the body.

The first law of thermodynamics says that, if the internal energy U of a body changes by an amount dU, then this change is equal to the sum of energy transferred as heat, and energy transferred as work.

$$dU = đQ + đW. \tag{27.1}$$

The Clausius formulation of the second law of thermodynamics says that there can be no process whose sole net effect is to transfer heat from a cooler to a warmer body. It follows from the second law that any two reversible heat engines operating between heat reservoirs at given temperatures T_1 and T_2 have the same efficiency, and that no engine is more efficient than a reversible one. Considerations of the Carnot cycle lead to the conclusion that, if a system that goes through a reversible cycle that leaves it in the same thermodynamic state as it started,

$$\oint \frac{đQ}{T} = 0. \tag{27.2}$$

And from this, it follows that there is a function S of the thermodynamic state, such that, in any reversible process, the change in S is given by

$$\Delta S = \int \frac{đQ}{T}. \tag{27.3}$$

This function (defined only up to an additive constant), is called the *entropy*.

[5] For further discussion of the probabilistic turn in Boltzmann's thinking, see Uffink (2007), Brown et al. (2009).

As already mentioned, on the kinetic theory, we should expect that not all molecules in a gas will have the same velocity, and that, as the molecules bounce around, there will be differences in local averages of kinetic energy of molecules from place to place. Therefore, since, on the kinetic theory, the temperature of a gas is proportional to the mean kinetic energy of its molecules, temperature differences will arise via spontaneous fluctuations, without expenditure of work, in contradiction to the second law. We can also expect, however, that these fluctuations will for the most part be negligible on the macroscopic scale, and large fluctuations will be both rare and unpredictable. Thus, though the second law of thermodynamics, as originally conceived, is untenable, we can set ourselves the goal of recovering from statistical mechanics a weakened version, which nonetheless would explain the evidence that led to acceptance of the stronger version. What the Clausius version of the second law deems impossible, namely, the transfer of heat from a cooler to a warmer body unaccompanied by a compensating increase of entropy, the revised version declares to be highly improbable.

Maxwell would add a further limitation. Note that, in the quotation in the previous section, the improbability of reversal of the mixing of the balls is limited to circumstances in which there is no sorting of white from black. For Maxwell, the validity of even the weakened version is restricted to situations in which we are dealing with molecules in bulk and there is no manipulation of individual molecules (Maxwell 1871, pp. 328–9).[6]

On Maxwell's view, the distinction between heat and work is not inherent in a physical process but has to do, rather, with the means available to us to keep track of and manipulate the motion of molecules.

> Available energy is energy which we can direct into any desired channel. Dissipated energy is energy we cannot lay hold of and direct at pleasure, such as the energy of the confused agitation of molecules which we call heat. (Maxwell 1878a, p. 221; Niven 1890, p. 646)

To a being such as Maxwell's demon, able to track individual molecules, "the distinction between work and heat would vanish, for the communication of heat would be seen to be a communication of energy of the same kind as that which we call work" (Maxwell 1878b, p. 279; Niven 1890, p. 669). With the vanishing of distinction between heat and work also vanishes any possibility of formulating thermodynamics. In particular, since the very definition of thermodynamic entropy requires a distinction between heat and work, for Maxwell, the entropy change associated with a process will not be an intrinsic property of the process—though, one might add, because of the vast gulf in scale between the macroscopic and the level of individual molecules, for macroscopic phenomena the concepts of heat and work will be sufficiently unambiguous to admit of unproblematic application.[7]

If it is a revised version of the second law of thermodynamics that we aim to recover from statistical mechanics, one according to which the processes declared impossible by the original version of the second law are judged improbable, then, it seems, there will be no avoiding the use of probabilistic concepts in statistical mechanics. This renders questions

[6] This is what the creature now known as "Maxwell's demon" is meant to illustrate. The demon is first described in a letter dated December 11, 1867, from Maxwell to P. G. Tait (Knott, 1911, pp. 13–14), and makes its first public appearance in Maxwell's *Theory of Heat* (1871), in a section entitled, "Limitation of the Second Law of Thermodynamics."

[7] For further discussion of the Maxwellian view of thermodynamics and statistical mechanics, see Myrvold (2011).

about the status of probabilities in statistical mechanics central to the interpretation of the theory.

Probabilities play somewhat different roles in Boltzmannian and Gibbsian approaches to statistical mechanics. Both make use of the apparatus of phase space and Hamiltonian dynamics. In the next section, this apparatus will be briefly reviewed.

27.5 Basic Concepts of Hamiltonian Dynamics

The Hamiltonian formulation of classical mechanics has turned out to be a useful setting for classical statistical mechanics. Consider a system of N degrees of freedom, represented by coordinates $\{q_1,...,q_N\}$. These might, for example, be the $3n$ position coordinates of n point particles; they might also include angle variables or other parameters. With each generalized coordinate q_i is associated a *conjugate momentum* p_i.

Since the Newtonian equations of motion are second-order in the time derivative, to specify a solution it does not suffice to specify the values of coordinates at a given time. We can, however, specify a solution by specifying values of the coordinates and their rates of change, or, equivalently, by specifying the coordinates and momenta. The $2N$-dimensional space whose points are given by specifying the coordinates and momenta of a system with N degrees of freedom is called the *phase space* of the system.

The dynamics of the system are encoded in a function on phase space called the *Hamiltonian*, which, for the systems with which we will be concerned, is simply the total energy of the system, expressed in terms of generalized coordinates $\{q_1,...,q_N\}$ and their conjugate momenta $\{p_1,...,p_N\}$. The dynamically possible trajectories through phase space are those that satisfy *Hamilton's equations of motion*,

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i}. \tag{27.4}$$

These equations define a flow on phase space; there is a function T_t that maps the phase space into itself, such that, if x is the phase point at some time t_0, $T_t x$ is the phase point at time $t_0 + t$.

The phase space volume of a subset A of phase space is given by

$$m(A) = \int_A dq_1...dq_N\, dp_1...dp_N. \tag{27.5}$$

Note that this is defined in terms of canonical coordinates and momenta. It is invariant under canonical transformations, that is, coordinate transformations that preserve the equations of motion (27.4), but not under arbitrary coordinate transformations. In particular, it makes a difference whether we use momenta or velocities to parameterize the space. For example, if we consider a system confined to a finite volume that contains two molecules of different masses, then the set of all states in which the more massive molecule has its velocity within certain limits will have larger phase space volume than the set of states in which the less massive molecule has its velocity within the same limits, although, on the

measure corresponding to (27.5) with velocities in place of momenta, these sets would have equal measure.

For any subset A of a phase space Γ, let $T_t(A)$ be the set of points that evolve, in time t, from points in A.

$$T_t(A) = \{T_t\, x \,|\, x \in A\}. \tag{27.6}$$

It is easy to show that phase space volume is preserved under the dynamical evolution (27.4),

$$m(T_t(A)) = m(A). \tag{27.7}$$

A probability distribution P_0 over the state of the system at time t_0, together with the phase-space flow T_t, determines a probability distribution P_t for any other time $t_0 + t$:

$$P_t(x \in A) = P_0(x \in T_t^{-1}(A)). \tag{27.8}$$

If the probability distribution for the state at time t is represented by a density function $\rho(\mathbf{q}, \mathbf{p}, t)$, this will obey *Liouville's equation*:

$$\frac{\partial \rho}{\partial t} + \sum_{i=1}^{N} \left(\frac{\partial \rho}{\partial q_i} \frac{\partial H}{\partial p_i} - \frac{\partial \rho}{\partial p_i} \frac{\partial H}{\partial q_i} \right) = 0. \tag{27.9}$$

It follows from Liouville's equation that any probability distribution given by a density function that is a function of the Hamiltonian will be a stationary distribution.

27.6 Boltzmannian Statistical Mechanics

27.6.1 Entropy and Probability

As already mentioned, in 1872 Boltzmann proved (with implicit assumption of the *Stoßzahlansatz*) that molecular collisions in a gas would lead to the Maxwell-Boltzmann distribution of velocities. The proof proceeded by defining a quantity that Boltzmann called H and showing that it tends to decrease. For an ideal gas, at least, the negative of H is related to the thermodynamic entropy. Later (1877b), he showed that there is a relation between H and phase-space volume, suggesting that the entropy of a macrostate is related to its phase-space volume. On a probability measure that assigns probabilities to regions of phase space that are proportional to their phase-space volume, entropy is then connected with probability.

In this section we follow the procedure of Boltzmann (1877b, [1896, 1898] 1995), which is summarized in Ehrenfest and Ehrenfest (1912). For simplicity, we consider a system that consists of a large number N of identical molecules, each with r degrees of freedom (the generalization to systems consisting of several types of molecules is straightforward). Let μ be the $2r$-dimensional phase space of an individual molecule, and let $\Gamma = \mu^N$ be the $2rN$-dimensional phase space of the entire system of N molecules.

We will assume that we need consider only a finite region of the system's phase space. It might, for example, be a gas confined to a box, with energy known to lie within a small

interval $[E, E+\delta E]$. For each molecule, there will be an accessible region of its phase space, consisting of states consistent with the constraints on the system as a whole (every molecule will have its position in the box, and no molecule can have an energy greater than the energy of the whole system). Partition the accessible region of μ into small regions $\{\omega_i, i = 1,...,m\}$ of equal phase-space volume $[\omega]$, corresponding to small intervals of values of each of the coordinates and momenta. Suppose that the macrostate of the system depends only on the number of molecules whose phase-point lies in each region ω_i.[8] Let $\{n_i, i = 1,...,m\}$ be these occupation numbers, that is, a specification, for each ω_i, of the number of molecules whose state lies in that region; such a specification is called a *state distribution*. For each state-distribution Z there is a corresponding subset Γ_Z of Γ, consisting of phase points that yield that state distribution (such a region is called, by the Ehrenfests, a "Z-star"). Define a function H of state-distributions,

$$H(Z) = \sum_{i=1}^{m} n_i \log n_i. \tag{27.10}$$

This H is the quantity that Boltzmann had argued, in 1872, would be decreased by collisions between molecules in a gas, until it reached its minimum possible value. For large N, we have a relation between $H(Z)$ and the phase-space volume of the Z-star Γ_Z.

$$\log\left(m(\Gamma_Z)\right) \approx -H(Z) + C, \tag{27.11}$$

where C is a constant that depends on N and on the size of the cells in our chosen partition of the molecular phase-space μ.

The relation (27.11) reveals the significance of H as an indication (up to an arbitrary constant) of the volume of phase space occupied by a state-description. Moreover, if we take Z_{max} to be the state description that minimizes H (that is, maximizes $-H$), subject to the imposed constraints, then we find that, for an ideal gas, the quantity

$$S_B = -kH(Z_{max}) \tag{27.12}$$

is equal, up to an additive constant, to the thermodynamic entropy.

This suggests a construal of entropy in terms of phase-space volume. For any phase point x, let $Z(x)$ be the Z-star containing x, and define

$$S_B(x) = k\log[m(Z(x))]. \tag{27.13}$$

One can generalize this to situations in which the macrostate is not a function only of occupation numbers of regions of the single-particle space μ.[9] Suppose the macrostate of

[8] This is a nontrivial assumption, valid for an ideal gas, but not for systems for which intermolecular potentials make a nonnegligible contribution to the total energy. For such systems, the total energy is not a function only of occupation-numbers of a partition of the single-molecule phase space, but also on the distribution of *pairs* of molecules in the two-molecule phase space. See Jaynes (1965, 1971) for discussion. Jaynes' essential point is correct, though it is marred by his taking (27.13), rather than its generalization (27.14), as the definition of the Boltzmann entropy.

[9] It is this generalization that is referred to in current presentations of the Boltzmannian approach to statistical mechanics. See, e.g., Lebowitz (1993, 1999); Goldstein (2001).

the system is defined by the values of a small number of functions $\{X_1, ..., X_k\}$. Partition the accessible phase space Γ into regions corresponding to small intervals of values of these macrovariables; each such region consists of points that, for practical purposes, share the same values of the macrovariables. Then the entropy assigned to a point x is given by

$$S_B(x) = k \log[m(M(x))], \tag{27.14}$$

where $M(x)$ is the macrostate containing x.

This gives an appearance of assigning an entropy that is a property of the physical state of the system alone. But note that the value of the Boltzmann entropy depends, not only on the phase point x, but also on the macrovariables chosen to define macrostates (presumably, these are the ones that we are able to measure), and a partition of the macrovariables fine enough that differences within a set are regarded as negligible; this, presumably, has to do with the precision with which we can measure the macrovariables.

Given a partition of the accessible phase space into macrostates, we identify the equilibrium macrostate with the one that has largest phase-space volume. For an ideal gas, and some other systems, the ratio of this volume to the volume of all other macrostates will be of the order 10^N, where N is the number of molecules. If we identify macroscopic systems as those containing a number of molecules roughly on the order of Avogadro's number—that is, on the order of 10^{23}— the equilibrium macrostate will have vastly larger phase-space volume than the rest of the accessible region of phase space.

27.6.2 Explaining Entropy Increase

These considerations give intuitive content to the H-theorem. The move from a non-equilibrium macrostate to the equilibrium macrostate is a move from a region that occupies a vanishingly small volume of the accessible phase space to a region that occupies most (as measured by phase-space volume) of the accessible region of phase space.

Can such considerations lead to a conclusion that, for a macroscopic system in a non-equilibrium macrostate, the system will, with overwhelming probability, relax to equilibrium?

Note that, even if we take the uniform measure on phase space to yield a probability measure, the observation that the equilibrium macrostate dominates the phase space does not suffice for the conclusion. What is required is that, for any non-equilibrium macrostate M, most of the states in M are ones that evolve into the equilibrium macrostate. This will be the case if the dynamics are ergodic (see §27.8.2.1), though ergodicity is not required for the conclusion to go through.

Sheldon Goldstein argues that we should expect most phase points in any non-equilibrium macrostate to move into the equilibrium state, on the grounds that

> [f]or a nonequilibrium phase point X of energy E, the Hamiltonian dynamics governing the motion X_t arising from X would have to be ridiculously special to avoid reasonably quickly carrying X_t into Γ_{eq} and keeping it there an extremely long time — unless, of course, X itself were ridiculously special. (Goldstein 2001, p. 43)

Note, also, that the argument, though it makes reference to the uniform measure on phase space, does not depend sensitively on this measure being regarded as the one we use to

judge some initial conditions improbable or "ridiculously special." What is required is that, whatever measure we use to judge probabilities, it agrees with the uniform measure on which sets have small probability.

The argument, as it stands, is symmetric under time-reversal. It supports equally well the conclusion that, with the exception of ridiculously special states, the states in a non-equilibrium macrostate are those that evolved from an equilibrium macrostate a short time before. Yet, if we run across, say, a thermos bottle that happens to contain warm water and some ice cubes, we don't conclude that this condition probably arose from a state of uniform temperature a short while ago.

This leads us to ask what grounds we have for regarding the exceptional states, that give rise to anti-thermodynamic behaviour, as ridiculously special, with an attendant inference to ridiculously improbable. Indiscriminate application of this sort of reasoning would lead one to regard all out-of-equilibrium states as ridiculously special. Yet systems that are far from thermodynamic equilibrium are not (apparently) rare; they are seemingly ubiquitous. Our experience hardly lends support to the claim that out-of-equilibrium systems are atypical!

Of course, it is possible that our experience is misleading. One can imagine scenarios on which what we see is not even close to a fair sample of all that there is, and everything we see is atypical indeed. One such scenario is the Boltzmann-Schuetz cosmology, on which the Universe consists of a vast sea of matter whose overall state is thermal equilibrium, with occasional fluctuations here and there away from equilibrium (Boltzmann 1895, Boltzmann [1896, 1898] 1995, §90). Though they would be mind-bogglingly rare, there would also be low-entropy regions as large as the observable universe. On such a scenario, the states we see around us would not be typical states, as the very existence of living, experiencing beings requires low-entropy matter. One can, without contradiction, maintain that features that are ubiquitous in our experience are rare in the universe.

There is a consequence of this cosmology, however, that Boltzmann seems not to have noticed. On such a scenario, the vast majority of occurrences of a given non-maximal level of entropy would be near a local entropy minimum, and so one should regard it as overwhelmingly probable that, even given our current experience, entropy increases towards the past as well as the future, and everything that seems to be a record of a lower entropy past is itself the product of a random fluctuation. Moreover, on such a scenario you should take yourself to be whatever the minimal physical system is that is capable of supporting experiences like yours, and you should regard your apparent experiences of being surrounded by an abundance of low-entropy matter as illusory. That is, you should take yourself to be what has been called a "Boltzmann brain."[10]

This is a logically possible scenario. But not only does it involve rejecting judgments of what is typical that are based on experience (which tells us that out-of-equilibrium systems are ubiquitous), it even goes so far as to lead us to reject everything we experience as illusory. Empirical evidence does not support this cosmology. Yet it is physics that brought us to these considerations, physics based on empirical evidence that the world is to be described, at least approximately, as a large number of molecules evolving according to Hamiltonian dynamics. A theory that tells us that the experiments on which it is founded are illusory undermines its own empirical base.

[10] The term is due to Andreas Albrecht. It first appears in print in Albrecht and Sorbo (2004).

The conclusion to be drawn is that, whatever judgments may be warranted about probabilities of states of things, they are not to be based on considerations of phase-space volume alone.

27.7 Gibbsian Statistical Mechanics

The Gibbsian approach involves consideration of probability measures on the phase space of a system. Gibbs thought of probability in frequentist terms, and accordingly enjoined his readers to imagine a great number of independent systems of the same type, all with the same macroscopic properties, but different microstates. Thus, he referred to *ensembles* of systems, and thought of the probability assigned to a region A of phase space as closely approximating, for a sufficiently large ensemble of similarly-prepared systems, the fraction of systems in the ensemble whose microstate is in A.

The goal of statistical mechanics is to identify properties of mechanical systems that are analogues of thermodynamic quantities, in the sense that one can demonstrate, on the basis of the laws of mechanics and appropriate probabilistic assumptions, that, with high probability, to a high degree of approximation these properties stand in relations analogous to those of thermodynamics. According to Gibbs,

> A very little study of the statistical properties of conservative systems of a finite number of degrees of freedom is sufficient to make it appear, more or less distinctly, that the general laws of thermodynamics are the limit toward which the exact laws of such systems approximate, when their number of degrees of freedom is indefinitely increased. (Gibbs 1902, p.166)

Gibbs gave names to ensembles of particular interest: the microcanonical, the canonical, and the grand canonical. The microcanonical ensemble is meant to be appropriate for an isolated system in equilibrium whose energy is known. The canonical ensemble is appropriate for a system in thermal contact with a heat bath of fixed temperature, which can exchange energy but not material with its environment, so that it contains a fixed number of molecules. In a grand ensemble, the number of molecules is not held fixed, as there might, for example, be chemical reactions taking place. A grand canonical ensemble is a grand ensemble in thermal contact with a heat bath.

Though Gibbs spoke of ensembles, in keeping with his frequentism about probability, in what follows we will speak of probability distributions, without commitment as to whether these are to be thought of in frequentist terms. For our purposes, we need only consider in detail the microcanonical and canonical distributions, as the key conceptual issues associated with Gibbsian equilibrium probability measures arise already with them. The reader should be aware, however, that the scope of statistical mechanics is not limited to considerations of systems with a fixed number of degrees of freedom.

27.7.1 Microcanonical Distributions

Consider a system whose total energy is known to lie within a small interval $[E, E + \delta E]$. Suppose also that the system is confined to a finite phase volume within this energy shell

(the system might, for example, be a gas confined to a box of finite volume). We define a phase-space measure that is uniform, in phase-space variables, within the accessible region of the energy shell, and zero outside of it. Since the density function is a function of energy alone, this is a stationary distribution.

If Γ is a $2N$-dimensional phase space of a system with N degrees of freedom, the subset Γ_E of all points having an energy E is a $2N-1$-dimensional surface within Γ. We can define, as a limiting case, a distribution on this surface, which will be the projection onto the energy surface of the uniform distribution in the energy shell.

27.7.2 Canonical Distributions

A canonical distribution is one given by a density function that takes the form

$$\rho(x) = Z^{-1} e^{-\beta H(x)} \tag{27.15}$$

for x in the accessible region of the system's phase space. Z is a normalization constant satisfying

$$Z = \int e^{-\beta H(x)} dx, \tag{27.16}$$

where the integral is taken over the accessible region of phase space. Z is a function of the parameter β, and any external parameters on which the accessible region of phase space or the Hamiltonian depend. It is known as the *partition function*.

Suppose we have two systems S_1, S_2 that are weakly coupled, so that the total Hamiltonian is approximately the sum of the Hamiltonians of the two systems. Suppose that the two systems initially are characterized by canonical distributions with parameters β_1 and β_2, respectively. Then the joint distribution will be an approximately stationary one if and only if $\beta_1 = \beta_2$. This, Gibbs argued, suggests that the canonical distribution is appropriate for representing a system in thermal equilibrium with a heat bath, with a temperature that is a function of β.[11] Considerations of the canonical ensemble applied to an ideal gas lead to the identification

$$\beta = \frac{1}{kT}, \tag{27.17}$$

where T is the absolute temperature and k is Boltzmann's constant.

27.7.3 The Gibbs Entropy

Gibbs argued that, for systems for which the canonical distribution is appropriate, the quantity,

$$S_G = -k\langle \log \rho \rangle = -k \int \rho(x) \log \rho(x)\, dx \tag{27.18}$$

[11] As Gibbs is careful to point out, this does not amount to a proof, as there are other distributions that share this property (Gibbs 1902, pp. 35–42).

behaves like the thermodynamic entropy (see Gibbs 1902, pp. 43–5 and ch. 14). Though Gibbs argument is restricted to situations for which the canonical distribution is appropriate, we can consider the quantity

$$S_G[\rho] = -k \int \rho(x) \log \rho(x) dx, \tag{27.19}$$

for other distributions. $S_G[\rho]$ is, in some sense, a measure of how "spread out" the probability distribution is.[12] This quantity has come to be known as the *Gibbs entropy* of the probability distribution given by the density function ρ. For any ρ, $S_G[\rho]$ is conserved under Hamiltonian flow.

27.7.3.1 *Gibbs entropy and Boltzmann entropy compared*

It can be shown that the standard deviation of energy

$$\Delta E = \sqrt{\langle E^2 \rangle - \langle E \rangle^2} \tag{27.20}$$

yielded by a canonical distribution will, for systems of very many degrees of freedom, be small compared to the expectation value of energy,

$$\frac{\Delta E}{\langle E \rangle} \sim \frac{1}{\sqrt{N}}. \tag{27.21}$$

Recall that, for macroscopic systems, N is on the order of Avogadro's number, that is, on the order of 10^{23}, so the deviation in energy is very small indeed. The energy is almost certain to depart only negligibly from its expectation value, and so the canonical distribution can be replaced, for the purpose of calculating expectation values of thermodynamic quantities, with a microcanonical distribution on the energy surface corresponding to the expectation value of energy.

Moreover, most of this energy surface will be occupied by the equilibrium macrostate, and there is little difference between calculating the phase-space volume of the energy surface and the volume of its largest macrostate. Thus, for systems in equilibrium and macroscopically many degrees of freedom, the Boltzmann entropy and the Gibbs entropy will be approximately equal, up to a constant, and, crucially, will exhibit the same dependence on external parameters.

Suppose we extend the identification of (27.19) as entropy for systems other than those in thermal contact with a heat bath. We might even extend this identification to non-equilibrium situations, for which thermodynamic entropy is undefined. Then, because of the measure-preserving property of Hamiltonian flow on phase space, for an isolated

[12] For a finite probability space, to the atoms of which are assigned probabilities $\{p_1, p_2, ..., p_n\}$, the Gibbs entropy becomes

$$S_G = -k \sum_{i=1}^{n} p_i \log p_i,$$

which is the quantity that Shannon (1948) named the *entropy* of the probability assignment $\{p_1, p_2, ..., p_n\}$, and for which he used the symbol H, in analogy with Boltzmann's H.

system, S_G will not increase with time. This makes it a poor candidate for tracking entropy changes in a process of relaxation to equilibrium. However, as suggested by Gibbs (1902, pp. 148–51), we can also define a coarse-grained entropy by partitioning the phase space Γ into small regions of equal volume, and replacing the probability distribution over microstates by one that is uniform over elements of the partition. The idea is that, if the elements of the partition are smaller than our ability to discriminate between microstates, this smeared probability distribution will yield virtually the same probabilities for outcomes of feasible measurements as the fine-grained distribution. The Gibbs entropy associated with this smeared probability distribution can increase with time.

Recall that the definition of the Boltzmann entropy also requires a coarse-graining of the phase space of the system. The conceptual differences between Boltzmann entropy and coarse-grained Gibbs entropy are not great.

27.8 Justifying Choice of Equilibrium Measures

The microcanonical distribution is uniform, in phase-space variables, within a small energy shell. One might be tempted to think that this is mandated by a straightforward application of the Principle of Indifference, and, indeed, some authors have taken some version of the Principle of Indifference as a fundamental postulate of statistical mechanics.[13] We should recall, however, the familiar fact that any application of a Principle of Indifference requires some judgment about which possibilities are equiprobable. In the case of a continuum of possibilities, as we have in classical statistical mechanics, an injunction to adopt a uniform probability measure requires specification of which variables the distribution is to be uniform in. A distribution uniform in canonical phase-space variables will not be uniform with respect to some other parameterization of the state space of the system. Even if we accept the authority of the Principle of Indifference, we ought to ask, why these variables, rather than some other parameters?

A further problem with adopting a Principle of Indifference in statistical mechanics is that there seems to be no reason for restricting its use to systems in equilibrium. One might be led by such a principle to adopt a probability distribution that is as uniform as possible, subject to compatibility with the current macrostate, even for systems that are far from equilibrium. But, as emphasized by Albert (2000), such a procedure would lead to disastrous retrodictions; a probability distribution of this sort would ascribe high probability to entropy increase both to the future and to the past of the current moment.

Part of the answer to the question about justifying the choice of measure lies in the fact that we are concerned with *equilibrium* measures. On the Gibbsian approach, thermal equilibrium is not to be thought of as a static state; it is one on which the microstate is constantly changing and the macrostate, though approximately constant most of the time, is subject to frequent tiny fluctuations and much rarer large ones. An ensemble of

[13] See e.g., Jackson (1968, pp. 8, 83) and Carroll (2010, pp. 166–9). E. T. Jayne's Principle of Maximum Entropy (Jaynes, 1957/1989a,b) is a version of the Principle of Indifference.

systems, however, should not exhibit any tendency to change overall, and this means that the equilibrium distributions should be *stationary* distributions.

As we have seen, it follows from the Liouville equation that, for a conservative system, any distribution given by a density function that is a function of the energy is a stationary distribution. It is thus easy to see that the microcanonical distribution is a stationary one. The question arises whether there might be other stationary distributions that are plausible candidates for an equilibrium probability distribution.

27.8.1 The Hypothesis of Uniform A Priori Probabilities

In an influential textbook published in 1938, Tolman introduced what he called "the fundamental hypothesis" of "equal a priori probabilities for equal regions in the phase space."

> Although we shall endeavour to show the reasonableness of this hypothesis, it must nevertheless be regarded as a postulate which can ultimately be justified only by the correspondence between the conclusions which it permits and the regularities in the behaviour of actual systems which are empirically found. (Tolman 1938, p. 59)

Tolman argues for the reasonableness of this postulate on the basis of Liouville's theorem, which shows that a distribution uniform in phase space is a stationary distribution; this shows that "the principles of mechanics do not themselves include any tendency for phase points to concentrate in particular regions of the phase space."

> Under the circumstances we then have no justification for proceeding in any manner other than that of assigning equal probabilities for a system to be in different equal regions of the phase space that correspond, to the same degree, with what knowledge we do have as to the actual state of the system. And, as already intimated, we shall, of course, find that the results which can then be calculated as to the properties and behaviour of systems do agree with empirical findings. (Tolman 1938, p. 61)

This is reminiscent of an invocation of a Principle of Indifference, albeit not an incautious one that ignores the necessity of a choice of variables over which to impose uniformity. It should be noted, however, that, for Tolman, the postulate is ultimately to be justified by empirical evidence.

27.8.2 Probabilities from Dynamics

27.8.2.1 *Approaches based on ergodic theory*

Boltzmann conjectured that

> The great irregularity of the thermal motion and the multitude of forces that act on a body make it probable that its atoms, due to the motion that we call heat, traverse all positions and velocities which are compatible with the principle of [conservation of] energy. (quoted in Uffink 2007, p. 40)

This (or rather, a variant of it) has come to be known as the *ergodic hypothesis*. As stated, it cannot be correct, as the trajectory is a one-dimensional smooth curve and so cannot fill

a space of more than one dimension. But it *can* be true that almost all trajectories eventually enter every open neighbourhood of every point on the energy surface. Boltzmann argued, on the basis of the ergodic hypothesis, that the long-run fraction of time that a system spends in a given subset of the energy surface is given by the measure that Gibbs was to call microcanonical.

Given a Hamiltonian dynamical system, and an initial point x_0, we can define, for any measurable set A such that the requisite limit exists, the quantity

$$\langle A, x_0 \rangle_{time} = \lim_{T \to \infty} \frac{1}{T} \int_0^T \chi_A(T_t(x_0))\, dt, \tag{27.22}$$

where χ_A is the indicator function for A,

$$\chi_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A. \end{cases} \tag{27.23}$$

$\langle A, x_0 \rangle_{time}$, provided it exists, is the fraction of time, in the long run, that a trajectory starting at the point x_0 spends in the set A.

A dynamical system is said to be *ergodic* if and only if, for any set A of positive measure, the set of initial points that never enter A has zero measure. It is easily shown that this condition is equivalent to *metric transitivity*: a dynamical system is *metrically transitive* iff, for any partition of Γ into disjoint subsets A_1, A_2 such that, for all t, $T_t(A_1) \subseteq A_1$ and $T_t(A_2) \subseteq A_2$, either $m(A_1) = 0$ or $m(A_2) = 0$.

Birkhoff (1931a,b) proved that, for any measure-preserving dynamical system, and any measurable set A,

1. The limit
$$\langle A, x_0 \rangle_{time} = \lim_{T \to \infty} \frac{1}{T} \int_0^T \chi_A(T_t(x_0))\, dt \tag{27.24}$$
exists for almost all points x_0. (That is, if X is the set of points for which this limit doesn't exist, $m(X) = 0$.)
2. If the dynamical system is ergodic, then
$$\langle A, x_0 \rangle_{time} = m(A). \tag{27.25}$$
for almost all x_0.

Consider an ergodic system that has been permitted to evolve in isolation for a long time. If we select a random time to look at it, then the probability that the system is in a subset A of Γ_E will be given by $\langle A, x_0 \rangle_{time}$, which, by Birkhoff's ergodic theorem, is equal to the measure ascribed to A by the microcanonical measure.

Is this a justification for taking the microcanonical measure to be the measure that yields the correct probabilities for an isolated system? Two reservations arise. The first has to do with whether actual systems of interest have ergodic dynamics. Proving this turns out to be very difficult even for some very simple systems. Moreover, there are systems, namely, those to which the KAM theorem applies, that are provably not ergodic.[14] The second is the use of

[14] See Berkovitz et al. (2006) for discussion of the applicability of ergodic theory.

the long-term time average. The picture invoked above, of a system isolated for a very long time and observed at a random time, does not fit neatly with laboratory procedures. One argument that has been given for considering the long-term time average is as follows.[15] Measurements of thermodynamic variables such as, say, temperature, are not instantaneous, but have a duration which, though short on human time scales, are long on the time scales of molecular evolution. What we measure, then, is in effect a time-average over a time period that counts as a very long time period on the relevant scale.

This rationale is problematic. The time scales of measurement, though long, are not long enough that the average over them necessarily approximates the limit in (27.22); as Sklar (1993, p. 176) points out, if they were, then the only measured values we would have for thermodynamic quantities would be equilibrium values. This, as Sklar puts it, is "patently false"; we are, in fact, able to track the approach to equilibrium by measuring changes in thermodynamic variables.

As mentioned above, if we are to ask for a probability distribution appropriate to thermodynamic equilibrium, the distribution should be a stationary distribution. The microcanonical distribution is a stationary distribution on Γ_E. If the system is ergodic, then it is the *only* stationary distribution among those that assign probability zero to the same sets that it does. For a justification of the use of the microcanonical distribution along these lines, see Malament and Zabell (1980).

27.8.2.2 *Almost-objective probabilities*

There is, in the mathematical literature on probability, a family of techniques that is known (somewhat misleadingly) as "the method of arbitrary functions." The idea is that, for certain systems, a wide range of probability distributions will be taken, via the dynamics of the system, into distributions that yield approximately the same probabilities for some statements about the system, because small uncertainties about initial conditions are amplified over time.[16]

It is plausible, at least, that the dynamics of the sorts of systems to which we successfully apply statistical mechanics exhibit the requisite sort of forgetting of initial conditions. Consider, for example, an isolated system that is initially out of equilibrium (it might, for example, be a cup of hot water with an ice cube in it). It is left alone to relax to equilibrium. Once it has done so, then, it seems, all trace of its former state has been lost, or rather, buried so deeply in the details of the system's microstate that no feasible measurement would be informative about it. For systems of this sort, a wide class of probability distributions over initial conditions evolves, via Liouville's equation, into distributions that, *as far as feasible measurements are concerned*, yield probabilities that are indistinguishable from those yielded by the equilibrium distribution.

We need not restrict ourselves to states of thermodynamic equilibrium. If we open a thermos bottle and find in it half-melted ice cubes in lukewarm water, it is plausible that

[15] Adapted from Khinchin (1949, pp. 44–5).

[16] The method of arbitrary functions was pioneered by von Kries (1886) and Poincaré (1912), and elaborated by a number of mathematicians, notably Hopf (1934, 1936). For a systematic overview of mathematical results, see Engel (1992); for the history, see von Plato (1983). See Myrvold (2012a,b) and Frigg (2016), (22) in this volume, for examples and discussion.

no feasible measurement on the system will determine whether the system was prepared a few minutes ago with only a little less ice, or an hour ago with boiling water and a lot of ice. If this is right, then again, a wide variety of probability distributions over initial conditions will evolve into ones that yield virtually the same probabilities for results of feasible measurements.

Ideas of this sort have recently drawn the attention of philosophers; see Strevens (2003, 2011), Rosenthal (2010, 2012), Abrams (2012), and Myrvold (2012a,b) for an array of recent approaches in which the method of arbitrary functions plays a role.

The method does not generate probabilities out of nothing; rather, the key idea is that a probability distribution over initial conditions is transformed, via the dynamical evolution of the system, into a probability distribution over conditions at a later time. Hence any use of the method must address the question: what is the status of the input distributions? Poincaré describes them as "conventions," which, it must be admitted, is less than helpful. Strevens (2003) is noncommittal on the interpretation of the input probabilities, whereas Strevens (2011) and Abrams (2012) opt for distributions based on actual frequencies.

Savage (1973) suggested that the input probabilities be given a subjectivist interpretation. For the right sorts of dynamics, widely differing subjective probabilities about initial conditions will lead to probability distributions about later states of affairs that agree closely on outcomes of feasible measurements; hence the output probabilities might be called "almost objective" probabilities. This suggestion is developed in Myrvold (2012a,b). The conception combines epistemic and physical considerations. The ingredients that go into the characterization of such probabilities are:

- a class $\mathcal{C}$ of credence-functions about states of affairs at time t_0 that is the class of credences that a reasonable agent could have, in light of information that is accessible to the agent,
- a dynamical map T_t that maps states at time t_0 to states at time $t_1 = t_0 + t$, inducing a map of probability distributions over states at time t_0 to distributions over states at t_1,
- a set $\mathcal{A}$ of propositions about states of affairs at time t_1, to which probabilities are to be assigned,
- a tolerance threshold ϵ for differences in probabilities below which we regard two probabilities as essentially the same.

Given these ingredients, we will say that a proposition $A \in \mathcal{A}$ has an *almost-objective probability*, or *epistemic chance*, if all probability functions in $\mathcal{C}$ yield, when evolved via T_t to t_1, essentially the same probability for A. That is, A has epistemic chance λ if, for all $P_0 \in \mathcal{C}$, $|P_t(A) - \lambda| < \epsilon$.

This concept includes an epistemic aspect, as an essential ingredient is the class $\mathcal{C}$ of credence-functions that represent reasonable degrees of belief for agents with our limitations.[17] This restriction would be eliminable if, for the events of interest, the dynamical map T_t yielded the same probabilities for absolutely all input measures, but this cannot be. Physics also plays a key role; the value of an epistemic chance, if it exists, is largely a matter of the dynamics.

Those who hold that epistemic considerations ought not to be brought into physics at all will not be happy with construing statistical mechanical probabilities in this way.

[17] Objective Bayesians would hold that this class is a singleton.

However, on the Maxwellian view of thermodynamics and statistical mechanics, on which the fundamental concepts of thermodynamics have to do with our ability to keep track of and manipulate molecular motions, this sort of blending of epistemic and physical consideration is just what one would expect to find in statistical mechanics.

27.8.2.3 *Probabilities from quantum mechanics?*

So far, we have been considering classical statistical mechanics. However, our world is not classical; it is quantum. Most writers on the foundations of statistical mechanics have assumed, implicitly or explicitly, that the conceptual problems of classical statistical mechanics are to be solved in classical terms; classical statistical mechanics should be able to stand on its own two feet, as an autonomous science, albeit one that gets certain facts about the world, such as the specific heats of non-monatomic gases, wrong.

One argument for this might be that we successfully apply statistical mechanics to systems for which quantum effects are negligible. This is questionable. The classical trajectories through phase space that exhibit anti-thermodynamic behaviour are unstable under random perturbations. Albrecht and Phillips (2014) estimate the relevance of quantum uncertainty to stock examples such as coin flips and billiard-ball gases, and conclude that "all successful applications of probability to describe nature can be traced to quantum origins."

As emphasized by Albert (2000, ch. 7), if we consider isolated quantum systems, and assume the usual Schrödinger evolution to be valid at all times, then this leaves us in pretty much the same conceptual situation as in classical mechanics. The dynamics governing the wave-function are reversible; for any state that exhibits the expected thermodynamic behaviour there is a state that exhibits anti-thermodynamic behaviour. Moreover, the von Neumann entropy—the quantum analog of the Gibbs entropy—is conserved under dynamical evolution. Considering nonisolated systems only pushes the problems further out; the state of the system of interest plus a sufficiently large environment can be treated as an isolated system; there will be states of this larger system that lead to anti-thermodynamic behaviour in the subsystem of interest.

If, however, collapse of the wave-function is a genuinely chancy, dynamical process, then things are different.[18] For any initial state, there will be objective probabilities for any subsequent evolution of the system. Albert (1994, 2000) has argued that these probabilities suffice to do the job required of them in statistical mechanics.

This is indeed plausible, though we lack a rigorous proof. If this proposal is correct, we should expect that, on time scales expected of relaxation to equilibrium, the probability distribution yielded by the collapse dynamics approaches a distribution that is appropriately like the standard equilibrium distribution, where "'appropriately like" means that it yields approximately the same expectation values for measurable quantities. It is not to be expected that the equilibrium distribution be an exact limiting distribution for long time intervals. In fact, distributions that are stationary under the usual dynamics (quantum or classical) will *not* be strictly stationary under the stochastic evolution of dynamical collapse theories such as the Ghirardi-Rimini-Weber theory (GRW) or Continuous Spontaneous Localization theory (CSL), as energy is not conserved in these theories. However, energy increase

[18] See Bassi and Ghirardi (2003) and Ghirardi (2011) for overviews of dynamical collapse theories.

will be so small as to be under ordinary circumstances unobservable; Bassi and Ghirardi (2003, p. 310) estimate, for a macroscopic monatomic ideal gas, a temperature increase on the order of 10^{-15} Celsius degrees per year. Thus, it is possible for collapse dynamics to yield relaxation to something closely approximating a standard equilibrium distribution, followed by exceedingly slow warming.

27.9 Puzzles about Equilibrium Measures, and a Resolution

Do the standard equilibrium measures represent objective features of the physical world, or should they be thought of as degrees of belief that we ought to have about the microstate of the system, given knowledge of the parameters that define the system's thermodynamic state?

On either account, we have a puzzle. They are said to be introduced on the basis of our incomplete knowledge of the state of a system. This suggests an epistemic reading. Nevertheless, we generate from them predictions about expectation values and fluctuations around these expectation values, predictions that can be tested by experiment. This suggests an ontic reading; we are not, when we are performing experimental tests of these predictions, probing the system to learn about our beliefs. We seem to require *both* an epistemic and an ontic reading, in an inconsistent way.

Taken literally, the standard equilibrium measure, applied to an isolated system that has recently relaxed to equilibrium from a non-equilibrium macrostate, is problematic on either an epistemic or an ontic reading, as it ascribes high probability to the system's having been in an equilibrium macrostate for an exceedingly long time. Since this was not the case, the measure does not reflect an objective chance distribution, and, since we know it was not the case, it does not represent our epistemic state.

This puzzle is easily resolved if we observe that use of the standard measures does not require a commitment to their representing a correct probability measure over the state of the system, where a "correct" probability measure might be an either objective chance measure or a credence function that represents our state of knowledge about the system. Consider a system that is, at t_0, in an equilibrium state with a known energy, subject to some constraint (say, a gas confined by a partition to one half of a box). The constraint is removed, and the system evolves, isolated, to a new equilibrium, at time t_1. Suppose, now, we apply at t_1 the microcanonical distribution appropriate to the new equilibrium state. This will not be the evolute of our initial probability distribution. However: unless there is some feasible measurement on the system that could be performed at t_1 that will discriminate between the system's having been at t_0 in the same equilibrium state it is in at t_1, and the state it actually was in, the microcanonical distribution will yield virtually the same probabilities as the evolute of our initial probability distribution, and we can use the microcanonical distribution for purposes of calculation, as a surrogate for the evolute of our initial probability distribution.

It seems to be an empirical fact that, for many systems of interest, the current macrostate of the system is all that is relevant to predictions about future measurements.[19] Since a range

[19] Note that this is not true for retrodictions!

of macroscopically distinguishable initial conditions can evolve to the same macrostate, this means that, for such systems, a wide range of probability distributions over initial conditions will evolve into distributions that yield substantially the same probabilities for outcomes of future measurements, and conditions are ripe for the existence of epistemic chances, as outlined above.

As we have characterized them, physical considerations as well as epistemic considerations come into the definition of epistemic chances, and, as has already been remarked, their values, if they exist, are largely a matter of dynamics. Though they have an epistemic aspect to them, we cannot ascertain their values by consulting our own cognitive states. An agent might have good reason to believe that a proposition has an epistemic chance, without knowing what its value is, either because she doesn't know the exact dynamics of the system (she might be uncertain, for example, whether a roulette wheel is biased), or because (and this is the condition we are in, for typical systems in statistical mechanics), the computational task of forward-evolving a reasonable credence-function via the actual dynamics of the system would simply be beyond the computational resources available to her.

This means that the values of epistemic chances are things that we can be uncertain about and can have credences about. Moreover, propositions about the values of epistemic chances can be put to experimental test. Consider some proposition A about a system, having to do with the result of some measurement subsequent to a time t_1. Suppose that we have good reason to believe that there is a probability p^* that represents the probability assigned to A by *any* probability distribution that results from forward evolving, via the actual dynamics of the system, some reasonable credence-function about the state of the system at time t_0. We can entertain hypotheses of the form $p^* = p$, for various values of p. Suppose that our conditional credences satisfy

$$cr(A \mid p^* = p) = p. \tag{27.26}$$

This is an analogue of the Principal Principle. Just as the Principal Principle turns frequency data into evidence about chances, this analogue turns frequency data from repeated experiments into evidence on which we can update our credences about the value of p^*.[20]

Now consider the hypothesis,

> The value of p^* is that given by the microcanonical distribution corresponding to the macrostate of the system at t_1.

Note that this hypothesis may be true, even if the microcanonical distribution is ruled out as a candidate for reasonable credence about the state of the system at t_1. If it is true, this means that our knowledge of the state of the system at t_0 has become irrelevant to the outcome of the measurement performed at t_1. The hypothesis is testable, and its experimental verification will give us a justification for use of the microcanonical distribution to calculate probabilities that we assign to outcomes of the experiment.

Tolman's fundamental hypothesis of equal a priori probabilities should be replaced with one that says,

> The correct probabilities for results of feasible future experiments are those that are yielded by a probability distribution that is uniform over the region of phase space corresponding to a system's macrostate.

[20] See Myrvold (2012a,b) for further discussion.

Construing the standard equilibrium distributions this way, as surrogates for more complicated distributions that result from time-evolved credence-functions over earlier states of affairs, resolves the puzzles associated with them.

With regard to the first puzzle, though these probabilities enter into our considerations because of incomplete knowledge of the state of a system, the value of an epistemic chance, if there is one, depends on the dynamics of the system, and is, moreover, the sort of thing that we can formulate testable hypotheses about. The Tolman-inspired approach, on which a posit about probabilities introduced on epistemic grounds is vindicated by experiment, begins to seem less mysterious. What is vindicated, however, is not Tolman's postulate, but the modified version above.

Typically, there will be a temporal asymmetry in this sort of use of equilibrium distributions. Use of an equilibrium distribution to calculate probabilities for measurements at t_1 will be justified only if our past knowledge of the state of the system has been washed out by the evolution of the system, and has become irrelevant for the purpose of anticipating the results of future measurements. Nothing can make our knowledge of the macrostate of the system at time t_0 irrelevant for *retrodictions* about the state of the system at time t_0 or before.

The source of the asymmetry lies in asymmetry of epistemic access. We can have memories and records of past events, whereas for future events of the sort considered we can typically do no better than to use our knowledge of the current state of the system and evolve it forward. There is no reason for the class $\mathcal{C}$ invoked in the characterization of epistemic chances to be invariant under time-reversal, and, typically, it will not be.

A common objection to the introduction of epistemic considerations into statistical mechanics is that our ignorance of the exact state of the world surely cannot explain why systems behave as they do.[21] This is correct! The coffee in my cup does not cool because I am ignorant of its exact microstate.

A system behaves as it does because of its dynamics, together with initial conditions. Explanations of relaxation to equilibrium will have to involve an argument that the dynamics, together with initial conditions of the right type, yields that behaviour, plus an explanation of why the sorts of physical processes that give rise to the sorts of systems considered don't produce initial conditions of the wrong type (or rather, don't *reliably* produce initial conditions of the wrong type). There is a connection, however, between the epistemic considerations we have invoked, and what would be required of an explanation of relaxation to equilibrium. The processes that are responsible for relaxation to equilibrium are also the processes that are responsible for knowledge about the system's past condition of non-equilibrium becoming useless to the agent. Thus, an explanation of relaxation to equilibrium is likely to provide also an explanation of washing out of the relevance to the future of knowledge about the past. Moreover, an explanation of why no process reliably produces initial conditions that lead to anti-thermodynamic behaviour would also explain the reasonableness of credences that attach vanishingly small credence to such conditions. Our judgments about what sorts of processes occur in nature and our judgments about what sorts of credences are reasonable for well-informed agents are closely linked; if there were processes that could reliably prepare systems in states that lead to anti-thermodynamic

[21] For a particularly vivid expression of this point, see Albert (2000, pp. 54–65); see also Loewer (2001, p. 611).

behaviour, then it would not be unreasonable for an agent to attach non-negligible credence to the system having been prepared in such a state, and we would adjust our judgments about what are and are not reasonable credences accordingly.

27.10 Conclusion

Quantum probabilities, viewed as objective chances, sidestep the above-mentioned puzzles associated with statistical mechanical probabilities, which have to do with how to mesh our use of probability with deterministic, reversible, dynamics, as the dynamics of collapse theories are neither deterministic nor reversible.

Construal of the probabilities in statistical mechanics as epistemic chances also resolves the puzzles associated with them. Moreover, the blending of epistemic and physical consideration employed in their definition is appropriate for statistical mechanics, if the goal is to recover thermodynamics viewed in a Maxwellian light. This is achieved without sacrificing the autonomy of classical statistical mechanics.

Acknowledgments

This work was supported by the Social Sciences and Humanities Research Council of Canada.

References

Abrams, M. (2012) Mechanistic probability. *Synthese.* 187. pp. 343–75.

Albert, D. Z. (1994) The foundations of quantum mechanics and the approach to thermodynamic equilibrium. *The British Journal for the Philosophy of Science*. 45. pp. 669–77.

Albert, D. Z. (2000) *Time and Chance*. Cambridge, MA: Harvard University Press.

Albrecht, A. and Phillips, D. (2014) Origin of probabilities and their application to the multiverse. *Physical Review D*. 90. p. 1235.

Albrecht, A. and Sorbo, L. (2004) Can the universe afford inflation? *Physical Review D*. 70. 063528.

Bassi, A. and Ghirardi, G. C. (2003) Dynamical reduction models. *Physics Reports.* 379. pp. 257–426.

Berkovitz, J., Frigg, R., and Kronz, F. (2006) The ergodic hierarchy, randomness and Hamiltonian chaos. *Studies in History and Philosophy of Modern Physics.* 37. pp. 661–91.

Birkhoff, G. D. (1931a) Proof of a recurrence theorem for strongly transitive systems. *Proceedings of the National Academy of Sciences.* 17. pp. 650–55.

Birkhoff, G. D. (1931b) Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences.* 17. pp. 656–60.

Boltzmann, L. (1872). Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe*. 66. pp. 275–370. English translation in Boltzmann (1966).

Boltzmann, L. (1877a) Bemerkungen über einige Probleme der mechanische Wärmetheorie. *Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe*. 75. pp. 62–100. Reprinted in Boltzmann (1909), pp. 116–122.

Boltzmann, L. (1877b) Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung resp. dem Sätzen über das Wärmegleichgewicht. *Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissenschaftliche Classe*. 76. pp. 373–435. Reprinted in Boltzmann (1909), pp. 164–223.

Boltzmann, L. (1895) On certain questions of the theory of gases. *Nature*. 51. pp. 413–15.

Boltzmann, L. ([1896,1898] 1995) *Lectures on Gas Theory*. New York, NY: Dover Publications.

Boltzmann, L (1909) *Wissenschaftliche Abhandlung*. Leipzig: J. A. Barth.

Boltzmann, L (1966) Further studies on the thermal equilibrium of gas molecules. In Brush S. G. (ed.), *Kinetic Theory*, Vol. 2: *Irreversible Processes*, pp. 88–175. English translation of Boltzmann (1872).

Brown, H. R., Myrvold, W. and Uffink, J. (2009) Boltzmann's *H*-theorem, its discontents, and the birth of statistical mechanics. *Studies in History and Philosophy of Modern Physics*. 40. pp. 174–91.

Brush, S. G. (1976a) *The Kind of Motion We Call Heat*. Book 1. Amsterdam: North-Holland.

Brush, S. G. (1976b) *The Kind of Motion We Call Heat*. Book 2. Amsterdam: North-Holland.

Carroll, S. (2010) *From Eternity to Here*. New York: Dutton.

Clausius, R. (1857) Über die Art der Bewegung, welche wir Wärme nennen. *Annalen der Physik*. 100. pp. 353–80.

Clausius, R. (1966) The nature of the motion which we call heat. In Brush, S. G. (ed.), *Kinetic Theory*, Vol. I. *The nature of gases and heat*. Oxford: Pergamon Press. Translation of Clausius (1857).

Ehrenfest, P. and Ehrenfest, T. (1912) *The Conceptual Foundations of the Statistical Approach in Mechanics*. New York, NY: Dover Publications.

Engel, E. M. (1992) *A Road to Randomess in Physical Systems*. Berlin: Springer-Verlag.

Frigg, R. (2016) Chance and Determinism. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Garber, E., Brush, S. G., and Everitt, C. W. F. (eds.) (1995) *Maxwell on Heat and Statistical Mechanics: On "Avoiding All Personal Enquiries" of Molecules*. Bethlehem, PA: Lehigh University Press.

Ghirardi, G. C. (2011) Collapse theories. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/archives/win2011/entries/qm-collapse/. [Accessed 25 Aug 2015.]

Gibbs, J. W. (1875) On the equilibrium of heterogeneous substances. *Transactions of the Connecticut Academy of Arts and Sciences*. 3. pp. 108–248, 343–524. (Reprinted in Gibbs 1906/1961, pp. 55–353.)

Gibbs, J. W. (1902) *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. New York, NY: Charles Scribner's Sons.

Gibbs, J. W. (1906/1961) *The Scientific Papers of J. Willard Gibbs*. New York, NY: Dover.

Goldstein, S. (2001) Boltzmann's approach to statistical mechanics. In Bricmont, J., Dürr, D., Galavotti, M., Ghirardi, G., Petruccione, F. and Zanghì, N. (eds.) *Chance in Physics. Lecture Notes in Physics*. 574. pp. 39–54. Berlin: Springer.

Hacking, I. (1975) *The Emergence of Probability*. Cambridge: Cambridge University Press.
Hájek, A. (1997) 'Mises redux'—redux: Fifteen arguments against finite frequentism. *Erkenntnis*. 45. pp. 209–27.
Hájek, A. (2009) Fifteen arguments against hypothetical frequentism. *Erkenntnis*. 70. pp. 211–35.
Hopf, E. (1934) On causality, statistics, and probability. *Journal of Mathematics and Physics*. 13. pp. 51–102.
Hopf, E. (1936) Über die Bedeutung der willkürlichen Funktionen für die Wahrscheinlichkeitstheorie. *Jahresbericht des Deutschen Mathematiker-Vereinigung*. 46. pp. 179–95.
Jackson, E. A. (1968) *Equilibrium Statistical Mechanics*. New York, NY: Dover.
Jaynes, E. (1957/1989a) Information Theory and Statistical Mechanics. *Physical Review*. 106. 4. pp. 620–30. Reprinted in Jaynes, E. T. *Papers on Probability, Statistics, and Statistical Physics*. Dordrecht: Kluwer Academic Publishers.
Jaynes, E. (1957/1989b) Information theory and statistical mechanics, II. *Physical Review*. 108. pp. 171–90. Reprinted in Jaynes, E. T. *Papers on Probability, Statistics, and Statistical Physics*. Dordrecht: Kluwer Academic Publishers.
Jaynes, E. T. (1965) Gibbs vs Boltzmann entropies. *American Journal of Physics*. 33. pp. 391–8. Reprinted in Jaynes (1989. pp. 79–88).
Jaynes, E. T. (1971) Violation of Boltzmann's *H* theorem in real gases. *Physical Review A*. 4. pp. 747–50.
Jeffrey, R. (1992) Mises redux. In *Probability and the Art of Judgment*. pp. 192–202. Cambridge: Cambridge University Press.
Khinchin, A. I. (1949) *Mathematical Foundations of Statistical Mechanics*. New York, NY: Dover.
Knott, C. G. (1911) *Life and Scientific Work of Peter Guthrie Tait*. Cambridge: Cambridge University Press.
Krönig, A. (1856) Grundzüge einer Theorie der Gas. *Annalen der Physik*. 175. pp. 315–22.
Lebowitz, J. L. (1993) Boltzmann's entropy and time's arrow. *Physics Today*. 46 (September). pp. 3–8.
Lebowitz, J. L. (1999) Statistical mechanics: A selective review of two central issues. *Reviews of Modern Physics*. 71. pp. 346–57.
Loewer, B. (2001) Determinism and chance. *Studies in History and Philosophy of Modern Physics*. 32. pp. 609–20.
Malament, D. B. and Zabell, S. L. (1980) Why Gibbs phase averages work: The role of ergodic theory. *Philosophy of Science*. 47. pp. 39–49.
Maxwell, J. C. (1867/1890) On the dynamical theory of gases. *Philosophical Transactions of the Royal Society*. 157. pp. 49–88. Reprinted in Niven, W. D. (ed.) *The Scientific Papers of James Clerk Maxwell*. Vol. 2. pp. 26–78. Cambridge: Cambridge University Press.
Maxwell, J. C. (1871) *Theory of Heat*. London: Longmans, Green, and Co.
Maxwell, J. C. (1878/1890a) Diffusion. *Encyclopedia Brittanica* (9th ed.), Vol. 7, pp. 214–21. Reprinted in Niven, W. D. (ed.) *The Scientific Papers of James Clerk Maxwell*. pp. 625–46. Cambridge: Cambridge University Press.
Maxwell, J. C. (1878/1890b) Tait's "Thermodynamics." *Nature*. 17. 257–59, 278–80. Reprinted in Niven, W. D. (ed.) *The Scientific Papers of James Clerk Maxwell*. pp. 660–71. Cambridge: Cambridge University Press.
Myrvold, W. C. (2011) Statistical mechanics and thermodynamics: A Maxwellian view. *Studies in History and Philosophy of Modern Physics*. 42. pp. 237–43.

Myrvold, W. C. (2012a) Deterministic laws and epistemic chances. In Ben-Menahem, Y. and Hemmo, M. (eds.) *Probability in Physics*. pp. 73–85. Berlin, Heideiberg: Springer.

Myrvold, W. C. (2012b) Probabilities in statistical mechanics: What are they? [Online] Available from: http://philsci-archive.pitt.edu/9236/. [Accessed 24 Aug 2015.]

Niven, W. D. (ed.) (1890) *The Scientific Papers of James Clerk Maxwell*. Vol. 2. Cambridge: Cambridge University Press.

Poincaré, H. (1912) *Calcul des probabilités*. Paris: Gauthier-Villars.

Poisson, S.-D. (1837) *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile, précédées des règles générales du calcul des probabilités*. Paris: Bachelier, Imprimeur-Libraire.

Popper, K. R. (1957) The propensity interpretation of the calculus of probability, and the quantum theory. In Körner, S. (ed.) *Observation and Interpretation: A Symposium of Philosophers and Physicists*. pp. 65–70. London: Butterworths.

Popper, K. R. (1959) The propensity interpretation of probability. *The British Journal for the Philosophy of Science.* 37. pp. 25–42.

Rosenthal, J. (2010) The natural-range conception of probability. In Ernst, G. and Hüttemann, A. (eds.) *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*. pp. 71–91. Cambridge: Cambridge University Press.

Rosenthal, J. (2012) Probabilities as ratios of ranges in initial-state spaces. *Journal of Logic, Language, and Information.* 21. pp. 217–36.

Savage, L. J. (1973) Probability in science: A personalistic account. In Suppes, P. (ed.) *Logic Methodology, and Philosophy of Science IV*. pp. 417–28. Amsterdam: North-Holland.

Shannon, C. E. (1948) A mathematical theory of communication. *The Bell Systems Technical Journal.* 27. pp. 379–423, 623–56.

Sklar, L. (1993) *Physics and Chance*. Cambridge: Cambridge University Press.

Strevens, M. (2003) *Bigger than Chaos: Understanding Complexity through Probability*. Cambridge, MA: Harvard University Press.

Strevens, M. (2011) Probability out of determinism. In Beisbart, C. and Hartmann, S. (eds.) *Probabilities in Physics*. pp. 339–64. Oxford: Oxford University Press.

Thomson, W. (1874) Kinetic theory of the dissipation of energy. *Nature.* 9. pp. 441–4.

Tolman, R. C. (1938) *The Principles of Statistical Mechanics*. Oxford: Clarendon Press.

Uffink, J. (2007) Compendium of the foundations of statistical physics. In Butterfield, J. and Earman, J. (eds.) *Handbook of the Philosophy of Science: Philosophy of Physics*, pp. 924–1074. Amsterdam: North-Holland.

Uffink, J. (2011) Subjective probability and statistical physics. In Beisbart, S. and Hartmann, C. (eds.) *Probabilities in Physics.* pp. 25–49. Oxford: Oxford University Press.

von Kries, J. (1886) *Die Principien Der Wahrscheinlichkeitsrechnung: Eine Logische Untersuchung*. Frieburg: Mohr.

von Plato, J. (1983) The method of arbitrary functions. *The British Journal for the Philosophy of Science.* 34. pp. 37–47.

von Plato, J. (1994) *Creating Modern Probability*. Cambridge: Cambridge University Press.

CHAPTER 28

PROBABILITY IN BIOLOGY

The Case of Fitness

ROBERTA L. MILLSTEIN

28.1 Introduction

The biological sciences make extensive use of probabilities, whether evolutionary biology, ecology, molecular biology, developmental biology, genetics, physiology, paleobiology, medicine, neurobiology, etc., and whether as parts of formal theoretical models or as parts of less formal mechanistic models. Organisms just don't seem to behave in determinate and fully regular ways. In principle, then, the discussion of probability in biology is a massive enterprise.

For example, evolution is probabilistic in the production of new variations (e.g., mutation; see Merlin 2010), in the microevolutionary perpetuation of those variations through natural selection and random drift, and even at the level of macroevolution; associated with these probabilities are a number of colloquial meanings of chance, such as "chance as ignorance," "chance as coincidence," or "chance as contingency" (see Millstein 2011 for an overview of the concept of chance in evolution). Models in ecology are also thoroughly probabilistic, with most lower-level details generally eschewed in equations containing only a few simple parameters (Colyvan 2005). In neurobiology, many phenomena are probabilistic, such as whether an Na^+ channel will open or a synapse will be potentiated; often the probabilities involved are low ones (Craver 2007). And any student of medicine is aware of the ubiquity of probabilities, particularly where uncertainty is involved (Djulbegovic, Hozo, and Greenland 2011).

However, most of the discussion of probability in the philosophy of biology has focused on probability in evolution, and much of that has centered on the concept of "fitness" (followed closely, perhaps, by discussions of "random drift"). In part, that is a reflection of the philosophy of biology's focus on evolution more generally, a focus that is now shifting to embrace the other areas of biology. Thus, it is somewhat apologetically that I find myself

perpetuating rather than bucking the older trend.[1] My reasoning is this: first, by discussing the probabilistic nature of fitness, I will be able to discuss a wide-ranging, established literature; and second, because the literature is so well-established, it can serve as a guideline for assessing where discussions of probability in biology can go wrong and where they can go right. Indeed, my goal in this chapter is to show how discussions of fitness in the philosophy of biology have wandered very far off track, and to try to gently nudge them back on track, pointing out important insights that were missed along the way.

I begin with a discussion of "fitness" in the work of Charles Darwin, which will also serve to introduce several key themes which are echoed in the contemporary discussion. I then provide an overview of the account of fitness that has garnered the lion's share of discussions about fitness, the propensity interpretation of fitness, followed by some critiques that have been offered of it. Many of these critiques have focused on the mathematical aspects of fitness rather than fitness as a propensity per se; I proceed to show how re-focusing on several aspects of the propensity interpretation of probability more generally can help to address the concerns that have been raised. I conclude with some general lessons for thinking about probability in biology.

28.2 Background: Darwin and Fitness

Perhaps surprisingly, the precise term "fitness" is not used in anything like its current evolutionary context in any of the six editions of Darwin's *Origin of Species*. Of course, the phrase "survival of the fittest" *is* well known;[2] Darwin used the phrase, due to Herbert Spencer, at the urging of Alfred Russel Wallace. The phrase made its first appearance in Darwin's *The Variation of Animals and Plants under Domestication*, and the fifth edition of the *Origin* was the first edition to contain it. However, it is instructive to see how Darwin uses very similar terms, such as "fit," "fitting," or "fitted," from the very first edition of the *Origin*. His uses are far too numerous to list here, but here are some typical ones:

> Let it be borne in mind how infinitely complex and *close-fitting* are the mutual relations of all organic beings to each other and to their physical conditions of life
>
> (Darwin 1859: p. 80; emphasis added).

> Or, again, the wolves inhabiting a mountainous district, and those frequenting the lowlands, would naturally be forced to hunt different prey; and from the continued preservation of the individuals best *fitted* for the two sites, two varieties might slowly be formed
>
> (Darwin 1859: p. 91; emphasis added).

[1] A second apology: although I attempt to survey a fair bit of the literature on fitness, the topic alone is so huge that (for reasons of space) there will be papers that I do not cite and concerns that I do not discuss. I have instead sought to focus on what I take to be the most central and interesting issues, though no doubt some of those have been inadvertently left out as well.

[2] Well known, and somewhat notorious, as the phrase has caused no end of trouble (e.g., charges, albeit false charges, that the theory of natural selection is circular).

> I look at all the species of the same genus as having as certainly descended from the same progenitor, as have the two sexes of any one of the species. Consequently, whatever part of the structure of the common progenitor, or of its early descendants, became variable; variations of this part would, it is highly probable, be taken advantage of by natural and sexual selection, in order to *fit* the several species to their several places in the economy of nature, and likewise to *fit* the two sexes of the same species to each other, or to *fit* the males and females to different habits of life, or the males to struggle with other males for the possession of the females
>
> (Darwin 1859: pp. 157–8; emphasis added).

One important thing to notice about these quotes is that being fit is a *relational* term. That is, organisms can be fitted to other organisms or to their "physical conditions of life" (external environment). More specifically, they can be fitted to their places in the "economy of nature" (i.e., the roles they play in an ecosystem), to members of the opposite sex in the same species, to their "habits of life," or for combat with members of the same sex for access to mates. In this way, being "fit" in an evolutionary context differs from the colloquial usage of the term; an organism is only fit relative to some aspect of the environment or to other organisms, not fit in an absolute sense. This can be seen most clearly in the wolf quote. Neither variety of wolf is absolutely fitter than the other; rather, one is fitter for the lowlands and one is fitter for the mountainous region. Contemporary views of fitness may differ on many points, but not on this one; fitness is always relative to the environment and to other organisms.[3] (I suspect that Darwin was referring to organisms being fitted to members of *other* species, as mistletoe is to certain trees, birds, and insects, one of Darwin's examples. Later, we will have reason to consider whether fitness ought to be relative to organisms of the same species as well).

The discussion of Darwin's views on fitness thus far makes it sound as though fitness (and by implication, natural selection) is a solely a matter of organisms' abilities to survive, without consideration of their abilities to reproduce, and indeed, much of the *Origin* reads that way. In particular, Darwin devotes an entire chapter (Chapter 3) to discussing what he called the "struggle for existence"; the chapter is mostly about the struggle to survive. However, reproductive ability does get brief and occasional mention; two oft-quoted passages are

> I should premise that I use the term Struggle for Existence in a large and metaphorical sense, including dependence of one being on another, and including (which is more important) not only the life of the individual, but *success in leaving progeny*
>
> (1859: p. 62; emphasis added)

and

[3] We will see below that fitness can be construed as relational in another sense as well, in *comparing* the fitness of one organism (or type) to another, as in the claim "*X* is fitter than *Y*." Consider "weight" as an analogy. An object has the weight that it does in virtue of being in a particular gravitational field. That is one sense of "relational," where "relational" means *context-dependent*. However, one can also speak of an object as being "heavy," which implies that it weighs more than another object (in a specified gravitational field). This seems to be a slightly different sense of "relational": *comparative*. In discussing the relation of fitness to the environment and other organisms, I mean to invoke only the context-dependent sense. Thanks to Marcel Weber for clarification on this point.

> can we doubt (remembering that many more individuals are born than can possibly survive) that individuals having any advantage, however slight, over others, would have the best chance of surviving and of *procreating their kind*?
>
> (1859: p. 81; emphasis added).

In the *Descent of Man*, Darwin is a bit more clear about including reproduction as part of his explanation of how to distinguish natural selection from sexual selection (see Millstein 2012 for discussion), although even there, he slips occasionally.[4] In any case, in contemporary evolutionary biology, reproductive ability is generally considered to be part of fitness,[5] although sometimes *viability* and *fecundity* are considered separately.

Finally, notice that I have repeatedly referred to fitness in terms of survival *ability* and reproductive *ability*. This is because Darwin says, repeatedly, that individuals having any advantage over others would have the *best chance* of surviving and of procreating (see, e.g., the quotes in the preceding paragraph). That is, the fitter organisms – the organisms with "structures" or other characteristics that provide advantages[6] over other organisms – may not in fact be the organisms that have the greater success in surviving and/or reproducing. Thus, from the very beginning, fitness (and thus natural selection) was seen as a *chancy* affair (Beatty 1984, Richardson and Burian 1992, Millstein 2002), paving the way for a contemporary understanding of fitness as probabilistic.

Although a full history of the term "fitness" is beyond the scope of this chapter (but see, e.g., Gayon 1998 and Jackson manuscript), it is important to understand that when evolutionary theory became mathematized in the early 20th century in the form of population genetics (see Millstein and Skipper 2007 for a discussion), fitness became a parameter used in equations. In other words, it became important to *quantify* organisms' fitnesses, to be able to say how much fitter some organisms were than others and to be able to use different fitness values to predict the frequencies of different types in the next generation. This, too, will be important for understanding contemporary debates over fitness.

28.3 The Propensity Interpretation of Fitness

Contemporary philosophical discussions of fitness generally begin with Susan Mills and John Beatty's classic 1979 article, "The Propensity Interpretation of Fitness," though proper credit must also be given to the Robert Brandon's (1978) independently and simultaneously developed propensity account of fitness (or, as he terms it, "adaptedness"). By the time of Mills and Beatty's article, most biologists were defining fitness in terms of an organism's

[4] For example: "Sexual selection acts in a less rigorous manner than natural selection. The latter produces its effects by the life or death at all ages of the more or less successful individuals" (1871/1981 vol. I. pp. 278–9).

[5] See Gayon (1998) for a discussion of how contemporary evolutionary biology came to emphasize the reproductive aspect of fitness and the role that eugenics played in that discussion.

[6] On Darwin's view, these advantages were usually seen as "slight" or "small"; over time, natural selection would aggregate those small fitness differences into adaptations.

actual survival and reproductive success – in other words, the actual number of offspring produced by an organism. Let us call this the *actualist definition of fitness*. Claims that evolutionary theory is untestable had been made and refuted, but the actualist definition of fitness caused evolutionary explanations invoking fitness to be circular, Mills and Beatty argued. If we seek to explain why type *A* is leaving a greater number of offspring than type *B*, the purported explanation "because *A* is fitter than *B*" is circular if "*A* is fitter than *B*" amounts only to the claim that "*A* left more offspring than *B*," as it does on the actualist definition of fitness. Mills and Beatty sought to solve this problem of *explanatory circularity* with the *propensity interpretation of fitness*.[7]

However, explanatory circularity was not the only problem that Mills and Beatty found with the actualist definition; they also noted that there was a mismatch between the actualist definition – that is, biologists' stated definition of "fitness" – and biologists' *usage* of the term "fitness". Let us call this the *mismatch problem*. They used two examples to illustrate the mismatch problem. The first example draws on what has become an oft-cited scenario from Scriven (1959). A pair of identical twins are standing together in a forest; one is struck by lightning before reproducing, while the other emerges unscathed and is later able to reproduce successfully.[8] The actualist definition of fitness would have us say that the second twin is far fitter than the first, since the actual reproductive success of the second twin is far greater than that of the first twin (whose reproductive success is zero). Yet, Mills and Beatty urge, this is counter-intuitive. If the twins are phenotypically and genetically identical (which I think is the supposition of the example), how can it make sense to say that one is fitter than the other? No biologist would actually use the term fitness in that way. The second example is also hypothetical but is somewhat more realistic:

> Imagine two butterflies of the same species, which are phenotypically identical except that one (*C*) has color markings which camouflage it from its species' chief predator, while the second (*N*) does not have such markings and is hence more conspicuous. If *N* nevertheless happens to leave more offspring than *C*, we are committed on the definition of fitness under consideration to conclude that (1) both butterflies had the same degree of fitness before reaching maturity (i.e., zero fitness) and (2) in the end, *N* is fitter, since it left more offspring than *C*
>
> (Mills and Beatty 1979: p. 268, n. 2).

And surely, it cannot be that the more conspicuous butterfly is fitter than the camouflaged butterfly! No biologist would use the term fitness in that way. The actualist definition of fitness is inadequate; it is mismatched with biologists' usage of the term.

Mills and Beatty's solution to the explanatory circularity problem and the mismatch problem is quite elegant and hearkens back to Darwin's way of thinking. The essence of their account is that fitness refers not to an organism's *actual* survival and reproductive success, but rather to its *ability* (i.e., disposition, capability, or tendency) to survive and reproduce. Thus, the twins were equal in fitness, whereas the camouflaged butterfly was fitter than the conspicuous butterfly. Actual reproductive success may be an indicator of an organism's

[7] Mills and Beatty's main goal was *not* to try to address the supposed circularity of the phrase "the survival of the fittest," although they thought that it could; they wisely note that "[t]his catch-phrase is not an important feature of evolutionary theory" (Mills and Beatty 1979: p. 270, n. 5).

[8] This would not be an instance of selection – or drift, for that matter – since there are no differences between the twins, and both selection and drift require variation.

ability, but it is a defeasible indicator; it can happen that the less fit butterfly out-reproduces the fitter butterfly. As Darwin seems to suggest, the fitter organism has the *best chance* of out-surviving and/or out-reproducing the less fit, but it may not do so. In other words, an organism has a *probabilistic* ability to survive and reproduce, i.e., it has a *propensity* to survive and reproduce.[9]

What does this propensity consist in? When we speak of other complex dispositional properties, such as solubility of salt in water (i.e., the propensity of salt to dissolve in water), we refer the *physical properties* of the salt (its ionic crystalline character) in the presence of the appropriate triggering conditions and in the absence of any countervailing conditions. Similarly, the fitness of an organism consists in its having physical traits that lend themselves to survival and/or reproduction in a particular environment; again, the camouflaged butterfly in a similarly colored environment can serve as an example. For both the salt and the butterfly, the presence of the appropriate physical properties of the entity, together with the presence of certain triggering conditions in the absence of countervailing conditions, causes (probabilistically, in the case of fitness) the manifestation of the specified behavior.

Thus, not only does the propensity interpretation solve the mismatch problem by proposing a view of fitness that is consistent with biologists' usage of the term, but it also addresses the explanatory circularity problem. With the propensity interpretation, if we seek to explain why type *A* is leaving a greater number of offspring than type *B*, the explanation "because *A* is fitter than *B*" means that *A* has a greater propensity than *B* to survive and reproduce in the given environment, which means that the physical properties of *A* in its environment are what cause it to tend to have greater reproductive success than *B* (with its physical properties). The relative physical abilities can be determined by engineering optimality models or other examinations of the physical properties of the organism in its environment, and then be confirmed by measurements of actual descendant contributions (Mills and Beatty 1979: p. 277; see also Brandon and Beatty 1984 and Brandon 1990 on this point). Also, Mills and Beatty's account of fitness incorporates the features of Darwin's discussed above: fitness is relative to the environment, fitness involves both survival and reproduction, and fitness is a chancy ability.

The propensity interpretation of fitness quickly gained a number of prominent adherents in addition to Mills, Beatty, and Brandon, such as Richard Burian (1983), Elliott Sober (1984), Philip Kitcher (1987), Ernst Mayr (1988), and Robert Richardson and Burian (1992). Of course, there were critics, too (Rosenberg 1985, Rosenberg and Williams 1986, Hodge 1987). Anyone familiar with the philosophy of probability might think that the lion's share of the criticisms came from proponents of other interpretations of probability, e.g., frequentist interpretations or epistemic interpretations. In fact, although there has been some discussion of alternate interpretations of probability for fitness and analysis of the propensity interpretation's appropriateness for the concept of fitness (e.g., Waters 1986; Richardson and Burian 1992; Bouchard and Rosenberg 2004; Abrams 2006, 2007), there has been surprisingly little.[10] Indeed, I would say that the fact that there have been so

[9] The probabilistic nature of natural selection (and thus fitness) is distinct from probabilistic random drift; see Beatty (1984) and Millstein (2002) for discussion.

[10] Some of the options can be ruled out quickly. An epistemic interpretation of probability would not be explanatory in this context, i.e., it would not explain why a population changed in frequency from

few publications on interpretations of probability as compared to publications on the topic of fitness more generally is one indication that the discussion has gone in some strange directions.

One might also expect that there would be some discussion of whether Mills and Beatty's proposed propensities were "true" (indeterministic) propensities, akin to the propensities involved in the radioactive decay of atomic particles. Mills and Beatty's view seems to have been that they are (1979: p. 273). Indeed, there has been a fair bit of discussion about whether evolution is indeterministic, with perhaps more authors arguing that it is than arguing that it isn't. My own view on this is that one can make sense of both deterministic and indeterministic propensities (Millstein 2003a) and that there is reason to be agnostic on the question of evolutionary indeterminism, at least for the time being (Millstein 2003b; this piece contains a review of the literature on evolutionary indeterminism up to that point).

28.4 The Propensity Interpretation of Fitness Takes a Mathematical Turn

Perhaps the view of fitness as a propensity would have rested there were it not for mathematical population genetics models, referred to at the end of Section 28.2. Again, it is generally not enough to say that *A* is fitter than *B*; one must also say *how much* fitter *A* is than *B*: 50% fitter? Twice as fit? etc. Note also that whether an organism reproduces or fails to reproduce is often not what is most relevant for natural selection. Reproduction is not an all or nothing affair; it matters *how many* offspring an organism has. Consider, for example, an organism that has zero offspring, an organism that has one offspring, and an organism that has twelve offspring. The effect on the distributions of types in the next generation of having one offspring is more similar to having zero offspring than it is to having twelve offspring. Furthermore, as Mills and Beatty point out, an organism can have different propensities to have different numbers of offspring. Perhaps one organism has a 25% chance of having zero offspring, a 50% chance of having one offspring, and a 25% chance of having twelve offspring (again, based on its physical traits in the particular environment). How does its fitness compare to a second organism that has a 25% chance of having zero offspring, a 25% chance of having one offspring, a 25% chance of having eight offspring, and a 25% chance of having twelve offspring? What changes in gene frequencies would we expect in the next generation based on these propensity values?

Mills and Beatty, despairing of the possibility of comparing one such distribution of reproductive propensities to another, suggest that the fitness value for an *individual organism* in its environment (fitness_1) be an organism's expected number of offspring in that environment, where the expected number is the weighted sum of the possible offspring contributions. In the previous paragraph, the fitness_1 of the first organism is .25*0 + .5*1 +

one generation to the next. And a frequentist view that interpreted the probabilities in terms of *actual* frequencies would not solve the explanatory circularity problem. See Brandon (1990) and Richardson and Burian (1992) for a discussion of the problems with using a limit or long-run relative frequency interpretation for fitness.

.25*12 = 3.5 whereas the fitness$_1$ of the second organism is .25*0 + .25*1+ .25*8 + .25*12 = 5.25. In other words, the second organism is fitter than the first because its expected number of offspring is greater: 5.25 as compared to 3.5. Mills and Beatty point out that biologists also often want to talk about the fitness of *types* (genotypes or phenotypes), where the fitness "reflects the contribution of a particular gene or trait to the expected descendent contribution (i.e., the fitness$_1$) of possessors of the gene or trait" (1979: p. 276). However, organisms possessing the gene or the trait in question will vary in their other genes or traits, and these genes or traits will affect an organism's fitness; thus, they define fitness$_2$, the fitness of types, as the *average* expected number of offspring for a given genotype or phenotype (i.e., the average fitness$_1$). This allows us to "predict and explain the evolutionary fate of the genes or traits which correspond to the alternate types" (Mills and Beatty 1979: p. 276). Finally, since we are generally interested in predicting or explaining the success of one type as compared to others in the population, Mills and Beatty define *relative* fitness$_2$ of a type as the fitness$_2$ of that type divided by the type that has the highest fitness$_2$ in the population (so that the type with the highest fitness$_2$ will always have a relative fitness$_2$ of 1).[11]

28.4.1 Non-mathematical Fitness or Mathematical Fitness?

It is hard to know how to reconcile fitness as a propensity, based on the heritable physical characteristics of an organism (i.e., the interpretation of fitness that addressed the explanatory circularity problem and the mismatch problem) with fitness$_1$, fitness$_2$, and relative fitness$_2$. In particular, there seems to have been a shift from the *causes* of an organism's survival and reproductive success (its physical traits in relation to its environment) to the expected *outcomes* (the expected survival and reproductive success) themselves. Are these different notions really part and parcel of the same propensity interpretation of fitness? Mills and Beatty do not offer us much guidance except to note that "the fitness values assigned to organisms [i.e., the expected reproductive success values] are not literally propensity values, since they do not range from 0 to 1"; instead, the fitness of an organism is "a complex of its various reproductive propensities" (1979: p. 275). Some philosophers of biology (e.g., Matthen and Ariew 2002, Pigliucci and Kaplan 2006), referring to the non-mathematical notion as "vernacular" fitness and the mathematical notion as "predictive" fitness, suggest that the two cannot be reconciled (and that the former is causal whereas the latter is not). Other philosophers of biology simply equate the propensity interpretation of fitness with the mathematical formulation (e.g., Abrams 2007), more or less ignoring the non-mathematical formulation, which avoids the issue altogether. Bouchard (2011) argues that a more general account of fitness will not focus on reproductive success. Brandon (1990) seems to differentiate what fitness *is* (how it is defined, what it is ontologically, which is the non-mathematical formulation) from how it is *measured* (the latter being consistent with Mills and Beatty's references to fitness$_1$, fitness$_2$, and relative fitness$_2$ as fitness *values*, i.e., the mathematical formulation, of which there may be many), which suggests that it is mistaken to think of the mathematical formulation of fitness as a definition of fitness at all (or as the propensity interpretation in particular).

[11] Note that this is "relational" in the sense of "comparative," rather than context-dependent. (See note 3).

Sober (2001), however, argues that reflections on the mathematical formulation of fitness call the propensity interpretation into question (despite his earlier endorsement of it).

How should we sort through these widely varying reactions for dealing with what Sober (2001) has dubbed "the two faces of fitness"? And can the mathematical and non-mathematical aspects of fitness be reconciled after all? My own sympathies lie most closely with Brandon's response, and I think that some sort of reconciliation may in fact be possible; however, to address Sober's rejection of Brandon's solution, we must first explore some of the mathematical complexities of using expected reproductive success as a proxy for distributions of propensities to produce different numbers of offspring.

28.4.2 Opening Pandora's Box

Ten years after propounding the propensity interpretation of fitness, Mills (now Finsen) and Beatty famously peeked inside Pandora's box and returned as critics, "to acknowledge and reframe some old problems, and to introduce some additional difficulties" (Beatty and Finsen 1989: p. 17). For a detailed discussion of these problems, I refer the reader to Beatty and Finsen's paper; here I seek only to summarize them. The first problem has to do with the time scale of fitness, a problem that Mills and Beatty (1979) addressed briefly in a pair of footnotes. Recall that fitness_1 is the expected number of offspring which an organism will leave in a given environment. But suppose organism x has a greater expected number of offspring than organism y, whereas y has a greater expected number of descendants than organism x in the subsequent generations. (Perhaps, for example, x cannot provide good offspring care for so many organisms, so that its offspring do not themselves have the ability to produce many offspring). Which is fitter, x or y? Mills and Beatty suggested that we differentiate "between long term fitness and short term fitness–or between first generation fitness, second generation fitness, . . ., n-generation fitness" (1979: p. 275, n. 10). Thus, it might turn out that while x is fitter in the short term, y is fitter in the long term. One might even consider as long a scale as "expected time to extinction" (Beatty and Finsen 1989: p. 18). Thus, perhaps we cannot say that there is one propensity interpretation of fitness, but rather, a (very large) family of fitness propensities.

A second set of problems discussed by Beatty and Finsen arises from the choice to use expected reproductive success – one value – to represent a distribution of different propensities to leave different numbers of offspring, even though there are other aspects of a distribution, such as its variance or its skew (asymmetry), that expected value does not reflect.[12] Their discussion derives from that of Gillespie (1974). Beatty and Finsen show that two genotypes can have the same expected reproductive success (fitness_2), but if one of the distributions has lower variance, that genotype will probably (i.e., is mathematically more likely to) be more reproductively successful than the other. Furthermore, if two genotypes have the same fitness_2 and their distributions have the same variance, but one has greater skew than the other, the one with greater skew will probably have greater reproductive success. Finally, in some cases, the greater skew of a distribution can compensate for

[12] Again, I refer the reader to Beatty and Finsen's article for the technical details behind the discussion in this paragraph and the next, details that are a matter of mathematics rather than biology and which are uncontested as far as I know.

larger variance, so that the genotype with greater skew but larger variance in its propensity distribution will probably have greater reproductive success. Thus, in all three of these cases, expected value yields the wrong prediction. In order to have a predictive calculation of future type frequencies for such cases, one must use the geometric mean rather than the arithmetic mean (i.e., the expected value). To make matters worse, whether we need to use the arithmetic mean or the geometric mean depends on whether there is variance within generations or between generations, respectively. But determining the distribution of propensities is hard enough; how will we determine whether the variance of the distribution manifests itself within generations or between generations, or even worse, whether organisms are using different strategies at different points in time? This is a problem, Beatty and Finsen assert, because "we may sometimes have no access to the sort of information we need in order to decide what statistic on the fitness distribution to employ in order to explain a particular evolutionary phenomenon" (1989: p. 28).

Beatty and Finsen suggest again that it seems as though there is a large family of propensities, rather than a single propensity. And then they conclude:

> It would be unfair to suggest that, lacking any generally agreed-upon definition of 'fitness', we therefore lack *any* understanding of evolution in terms of fitness differences and natural selection. On the other hand until we have an appropriate general definition of fitness, it is not altogether clear *how much* we understand about evolution in terms of fitness differences and natural selection
>
> (1989: p. 29; emphasis in original).

In short, Beatty and Finsen don't seem satisfied with a family of propensities of fitness; they suggest that greater understanding will be obtained only when we have a more general definition.

As with the original Mills and Beatty article, Beatty and Finsen's article has been quite influential, with a variety of conclusions being drawn, mostly arguing that the complications of variance and skew spell varying degrees of trouble for the propensity interpretation of fitness. Sober's response is particularly notable. Sober (2001), also drawing on the work of Gillespie (1974, 1977), argues that within-generation variance raises an additional problem, one that casts doubt on the propensity interpretation altogether. Sober points out that in order to make a predictive calculation of future frequencies when there is within-generation variance, one must include the *size of the population* in the calculation. But this is strange because the size of the population is not causally affecting organisms' reproductive success (or so Sober claims – I will challenge this claim below). Thus, Sober concludes, "an organism's fitness is *not a propensity that it has*—at least not when fitness must reflect the existence of within-generation variance in offspring number" (2001: p. 320; emphasis added). Rather, Sober suggests, fitness is a "holistic" quantity that reflects both properties that affect an organism's reproductive success (the organism's traits in relation to its environment) and properties that do not (the size of the population).

At this point the reader may be forgiven for being confused. Short-term vs long-term reproductive success? Between-generation variance? Within-generation variance? Skew? What does any of this have to do with solving the explanatory circularity problem? What does this have to do with solving the mismatch problem? What do these problems of making mathematical predictions for future generations have to do with whether fitness is a physical

property of organisms in their environments to survive and reproduce? In other words, what do these mathematical issues have to do with *the question of whether fitness ought to be construed as a propensity*? Are they even problems with propensities in particular? After all, *any* interpretation of probability is consistent with the use of probability distributions and their moments (a technical term referring to properties of a distribution like expected value, variance, and skew); under *any* interpretation of probability, the same problems with making predictions from probability distributions would still be present. It seems as though the debate has gotten off track. Perhaps a re-examination of the nature of propensities will bring us back on track.

28.5 Reflections on Propensities: Closing Pandora's Box?

Contemporary philosophers of probability often distinguish between two versions of the propensity interpretation, a long-run version and a single-case version:

> A long-run propensity theory is one in which propensities are associated with repeatable conditions, and are regarded as propensities to produce in a long series of repetitions of these conditions frequencies which are approximately equal to the probabilities. A single-case propensity theory is one in which propensities are regarded as propensities to produce a particular result on a specific occasion
>
> (Gillies 2000: p. 822; see also Gillies, this volume (19)).

Mills and Beatty do not distinguish between these two possibilities, and it is not clear to me that it is important that they do so. Would it help with the question of whether fitness refers to success after one generation, two generations, or $n>2$ generations, for example? Although some have tried to argue for an exclusively short-run or an exclusively long-run view of fitness, it seems clear that they are different and that our views of fitness must account for both. It is also clear that the propensity for long-run reproductive success is not simply the summation of many identical short-run reproductive successes, since the traits that tend toward long-run reproductive success (e.g., adaptability to different conditions) may be different from those that tend toward single-generation reproductive success (e.g., high fecundity).

28.5.1 Propensity to *X*

What *is* important to note, however, is that either way, a propensity is a tendency to produce a particular sort of *behavior* (again, whether one single outcome or outcomes in the long-run). That is, a propensity is always a *propensity to exhibit a behavior X*. It makes no sense to talk of a propensity in the absence of specifying *X*. This may seem like a trivial point, but I think it will help us to understand the question of whether we should be talking about reproductive success after one generation or after some number *n* generations.

The ability of an organism to have *i* descendants after one generation describes one behavior. The ability of an organism to have *j* descendants after two generations describes another. The ability of an organism to have *k* descendants after some specified *n* generations describes a third behavior. These are, in fact, different propensities, because they describe abilities for different behaviors. The different behaviors may, or may not, correspond to different underlying physical traits that give rise to the ability.

Consider an analogy. Dara Torres has a propensity to swim fast. So does Janet Evans. Both were, at one point in time, U.S. swimmers over the age of 40 vying to compete in the 2012 Summer Olympics,[13] and both have won gold medals in previous Olympic competitions. Who is the faster swimmer? Well, Torres is a sprinter. In the shorter distances (e.g., 50m freestyle), she would likely prevail. But Evans, the long-distance swimmer, would likely beat Torres in the 1500m freestyle. So, should we be worried about whether "being a fast swimmer" is a propensity? No. There is no puzzle here. These are different abilities for different behaviors that in all likelihood have different underlying physical manifestations (e.g., perhaps more fast-twitch muscles for Torres, more slow-twitch muscles for Evans). In other words, the propensity to succeed in the 50m freestyle is different from the propensity to succeed in the 1500m freestyle. And the two propensities are equally legitimate.

So, which is the real propensity of fitness, the ability to have many descendants after one generation or the ability to have many descendants after *n* generations? They are both equally deserving of the name; they are different propensities and different abilities. As Beatty and Finsen suggest, fitness should be understood as a family of propensities to survive and reproduce. (Similarly, we can understand "Olympic swimmer" as a family of propensities to swim fast). But at the end of their article, Beatty and Finsen bemoan the lack of a general view. There is no need to bemoan the lack of a general view.[14] We should not expect one; indeed, one would not be desirable if it led us to think the same set of physical traits gave rise to the ability to have short-run reproductive success as gave rise to the ability to have long-run reproductive success for a given genotype of a given species. Again, this is not likely, and certainly not guaranteed always to be the case.

28.5.2 Propensities are Relational

Another important aspect of propensities is one that Karl Popper emphasized. On Popper's view, propensities are not inherent in individual things, rather "they are *relational* properties of the experimental arrangement" (1959: p. 37; emphasis added). As we have already discussed, it is well agreed upon (since before Darwin and beyond) that fitness is relational. Darwin listed many such relations, whereas contemporary accounts of fitness, such as that of Mills and Beatty, specify only that fitness refers to the ability of an organism to survive and reproduce in its environment. But what does the environment include? In particular, does it include the population of which the organism is a member?

Sober asserts that the size of the population plays no causal role except in "special cases" such as density-dependent selection or frequency-dependent selection. He gives the

[13] Alas, neither succeeded in making the 2012 U.S. Olympic Swim Team.

[14] Sober (2001) also argues that there is no problem with accepting both a long-term concept of fitness and a short-term concept of fitness.

example of four cows standing in the four corners of a large pasture; two of them have within-generation variance in their associated probability distributions and two don't. Sober states, "The cows are causally isolated from each other, but the fitnesses of the two strategies reflect population size" (2001: p. 32). Recall that this is supposed to be a problem for the propensity account of fitness – a serious enough problem that implies we ought to abandon the propensity interpretation of fitness altogether. However, there are two problems with this example and its underlying assumptions. First, cows alone cannot form a population; one needs bulls.[15] I don't mean to be snarky here, but this matters. Whether the cows are choosing their mates or whether the farmers are choosing their mates for them, who is mated with whom and how many mates are available will affect each cow's reproductive success (i.e., their fitnesses). Second, it is simply implausible that each cow stays in her own corner, never venturing to eat another cow's food, drink her water, or occupy her shelter, which again would affect the reproductive success of the other cows. In short, as I have argued elsewhere, what makes a population a population at all is that there are causal interactions – in particular survival and reproductive interactions – between the organisms (Millstein 2009, 2010); if the cows really are not interacting, then they are not part of the same population and not undergoing the same selection process.[16] It would then not make sense to compare their fitnesses. Or, to put the point in a more Darwinian vein, the struggle for existence, which is a result of various checks on the population size (due to limited resources, climate, prey, etc.), is part of the process of natural selection. Without it, organisms could reproduce without bounds. Sober's example, even if it were possible, is not an example of natural selection.[17]

In short, fitness is an organism's propensity to survive and reproduce in a given environment and in a given population (or, more precisely, is a family of propensities to do so) – and not just in "special cases" of density-dependence and frequency-dependence. In other words, an organism's fitness is relative to both its environment *and* its population. (If that was always meant to be implicit, so be it.[18] The problem with leaving things implicit is that they can be forgotten). This suggestion is similar, although perhaps not identical, to that of Evelyn Fox Keller, who states,

> The fitness of a particular female (or male) reproducing sexually always depends, first, on the availability, and second, on the fertility of males (or females) in the breeding group in which that organism finds itself,

concluding that

> [f]or sexually reproducing organisms, fitness is in general not an individual property but a composite of the entire interbreeding population.
>
> (1987: p. 393)

[15] Thanks to a certain Iowa boy for pointing out what ought to be obvious to us city folk, but isn't always.

[16] Ariew and Ernst (2009) and Walsh (2007) also overlook this point.

[17] See Lennox and Wilson (1994) for an argument for why "struggle for existence" is a necessary condition for natural selection, and why changes due *only* to differences in intrinsic rate of reproduction ought to be considered a different evolutionary process. Note that Ariew and Lewontin's (2004) claim that the "two faces of fitness" cannot be reconciled relies on conflating these two processes into one.

[18] See Mills and Beatty (1979: p. 271, n. 6).

I would generalize her reasoning to include all the ways that organisms affect each others' survival – both positively and negatively – in the struggle for existence.

If this is right, then it is not puzzling that a calculation for the future relative frequencies in a population should contain population size as a parameter. Features of the population, including its population size, are causal factors in natural selection.[19] Thus, contra Sober, the inclusion of population size in certain calculations is no reason to reject fitness as (a family of) propensities.

28.5.3 Determining the Value of a Distribution of Propensities

Even if the arguments of Section 28.5.2 are accepted, it might seem as though Pandora's Box is still open; we still do not have a univocal way of calculating fitness values and thus no univocal way of saying whether one organism is fitter than another. Recall that the problem arises because for each organism, we can (in principle – this is hard to do in practice!) describe a distribution of propensities to have varying numbers of offspring. Using these probability distributions, we can easily calculate the expected reproductive success for each organism, the expected reproductive success for each type (by averaging all the organisms of that type), and the relative fitness of each type. However, in some cases, the expected value will give us a misleading result; *A* will have a higher expected reproductive success than *B*, but mathematically, *B* is likely to have a greater number of offspring than *A*. Again, as discussed above, the amount of variance or skew in a distribution affects the predicted number of descendants in future generations.

The basic problem, of course, is that we are trying to use one value to describe an entire distribution that contains many different characteristics – more technically, moments. Or as Beatty and Finsen put the point, "identifying fitness generally with any one statistic, or any particular function of statistics, is mistaken" (1989: p. 29). Each of those moments, whether expected value, variance, skew, or any other moment of the distribution, captures some aspect of the distribution while overlooking others. What we really want is to be able to compare one entire probability distribution to another, but there does not seem to be a way to do this. Or is there?

Recent papers by Sean Rice and colleagues (Rice 2008; Rice and Papadopoulos 2009; Rice, Papadopoulos, and Harting 2011) suggest that there might be.[20] In their model, an individual's fitness is the number of descendants that it has after some chosen time interval; they use descendants rather than offspring in order to allow their model to go beyond one generation (and we have already seen reasons for doing so). However, each organism is not assigned one particular value; rather, fitness is treated as having a distribution of possible values, each with a probability assigned to it. In each generation, of course, an organism will have a determinate number of descendants. Again, I refer the reader to the authors' papers for the technical details, but the basic idea is this. If we were to run a computer simulation of a population of organisms, each with a probability distribution of possible

[19] This is also to be expected if one understands natural selection to be a form of discriminate sampling; see Beatty (1984) and Millstein (2002). Sampling processes, whether discriminate or indiscriminate, are always affected by population size.

[20] Thanks to John Beatty for the suggestion to look at Rice's work.

offspring values, we could have the computer assign each organism an actual number of offspring in each generation based on the organism's probability distribution.[21] Offspring would be assigned a phenotype (and thus a distribution of *their* possible offspring values) based on a distribution of possible offspring phenotypes assigned to parents. With many runs of the simulation, we can thus see how different distributions compare, i.e., whether *A*, with its entire distribution, is in fact fitter than *B* by seeing whether it produces a greater number of descendants in most of the simulations, for as many generations as we care to examine.

I think the approach of Rice and colleagues is extremely promising; in addition to potentially addressing the worries raised by Beatty and Finsen, it has at least three other benefits. First, they are offering not just a way to understand fitness values; they are also offering a new way to model selection processes. More traditional approaches, such as those based on the Hardy-Weinberg equation or the Price equation, tend to treat fitness as a fixed parameter. Thus, there has always been a bit of a disconnect between the way we have understood the concept of fitness (as probabilistic) and the way that the models have treated fitness (as fixed and determinate). The model that Rice and colleagues propose explicitly incorporates a probabilistic notion of fitness, thus making our concept and our model more consistent with one another and yielding a model that is more realistic as well. Again, as Darwin himself noted, the fitter organism may not in fact be the one that is the more successful. A second benefit of their way of treating fitness is that, by considering the full distribution instead of a limited set of moments, new results have been found. For example, Rice (2008) characterizes new situations where the expected direction of evolution is toward the phenotype with *higher* variance in fitness. A third benefit is that situations where fitnesses are fluctuating due to fluctuating environment, or where the actions of organisms are affecting their fitnesses (niche construction), are more easily accommodated.

Perhaps most importantly, however, the approach of Rice and colleagues vindicates Brandon's (1990) suggestion that we should separate what fitness *is* from how we assign *values* to it. Expected reproductive success is not the propensity interpretation of fitness and it never was; it was just one way of trying to grapple with probability distributions and assign fitness values. Whether the particular approach of Rice and colleagues succeeds, or whether some variant is needed, shows that it is a *mathematical* problem as to how one compares probability distributions. Clearly, this is a difficult thing to do, but these difficult puzzles don't change the fact that it is the probability distributions which represent the value of an organism's fitness.

Fitness is an organism's propensity to survive and reproduce (based on its heritable physical traits) in a particular environment and a particular population over a specified number of generations.[22] That is what fitness *is*. A probability distribution of possible offspring contributions for particular organisms can be used to compare the fitness of organisms and to make predictions of future frequencies in the population. That is how

[21] Rice and colleagues also describe an analytical solution, but I think the computer model (they have performed Monte Carlo simulations) is a little easier to understand.

[22] Fitness is also arguably a property of other biological entities, such as genes, groups, and species; however, here I will set aside debates over which entities are proper units of selection and in what sense.

fitness can be assigned a *value*.[23] As for the best way to compare probability distributions, I leave that to mathematicians and mathematical biologists.

28.5.4 Propensities are Causal

Propensities are usually understood to be causal; indeed, that they are causal is one of the features of propensities that philosophers of probability find challenging, because their causal nature implies that they must *not* obey the usual (Kolmogorov) probability calculus (Hájek 2012).[24] That is, propensities are usually held to be the cause of observed, actual frequencies; the propensity of a fair coin to land heads 50% of the time is the cause of the actual sequence of heads and tails (e.g., HTTHHHTHHT). Yet Sober – in the same book in which he endorsed the propensity interpretation – famously argued that fitness is "causally inert" (unlike some propensities, which he thinks are causal) even though natural selection is one of the causes of evolution (Sober 1984; see also Hodge 1987). On the other hand, Brandon and Beatty state that "[a]ccording to the propensity interpretation, the connection between the ability and the actual manifestation of the ability is a causal connection" (1984: p. 344). Who is right?

Sober denies that "overall fitness is causally efficacious" (1984: p. 89). The word "overall" is key here. The essence of his argument is this: the overall fitness of an organism is made up of many different abilities (in relation to its environment), such as the ability to avoid disease and the ability to escape predation. However, Sober argues, in a given survival event, both of those abilities may not be relevant:

> When an individual survives because it manages to avoid being caught by a predator, I can see how its *invulnerability to predation* was a causal factor contributing to its survival. Likewise, when an individual is infected by a disease but survives because of its immunity, I can see how its *invulnerability to disease* helped keep it alive. But an organism's overall fitness–its high probability of surviving, no matter what cause of mortality may present itself–strikes me as being causally inert.
>
> (Sober 1984: p. 91)

The overall fitness of an organism is, Sober asserts, a disjunction of abilities. One of them may be causally efficacious in a particular case, but the disjunction is not. If an organism escapes a predator, its immunity to disease played no causal role,[25] and so, Sober suggests, it does not make sense to say that its overall fitness played a causal role.

However, it's not clear that "overall fitness" is the concept of fitness at stake; rather, it is arguably *trait fitness* that we ought to be considering.[26] Mills and Beatty state that biologists are generally talking about the relative fitness of types (genotypes or phenotypes) where

[23] This solution, if sound, might make Sober's worries about incorporation of population size into our calculations moot. However, the reasons for considering the fitness of an organism to be relative to its population (as well as to its environment) still hold.

[24] A worry which I will not be addressing in this chapter. The reader should note, however, that alternative probability calculuses have been developed.

[25] Note that to say that the traits of an organism can be separated causally is a substantive empirical assumption, given the high degree of integration that organisms have.

[26] See also Ramsey (2013) for a discussion of trait fitnesses in the context of Sober's arguments.

organisms share a gene or a trait but vary in other respects; what biologists are trying to do is to "explain the evolution and/or persistence of a gene or its phenotypic manifestation in a temporally extended population" by showing "that possessors of the gene or trait were generally better able to survive and reproduce than possessors of alternate traits or genes" (1979: p. 346). Indeed, many of the classic examples of natural selection focus not on an organism's fitness, but on the fitness of organisms given a certain specified trait: the beak of a finch, the color of a moth, the banding pattern of a snail. The other traits are backgrounded and assumed not to be correlated with the trait in question. To elaborate further, suppose that (as a defeasible hypothesis), a population had some organisms with trait *X* and other organisms with a different trait *Y*, but that either 1) other traits were distributed randomly across those two subgroupings, or 2) there was no difference in the distributions of other traits across the two subgroupings. These situations would be analogous to a randomized experiment or an experiment with matching, respectively. Under these conditions, trait fitness can be shown to be a causal factor contributing to organisms' survival in the population.

Note the focus on *relative* fitness (in the comparative sense) in the explanatory practices of biologists. Indeed, it *is* a mistake to say that the fitness of individual organisms and traits ($fitness_1$ and $fitness_2$, respectively) play a causal role *in natural selection*. Only relative fitness ($relative\ fitness_2$) matters for natural selection. The ability of a type to produce 100 offspring will be selectively favored only if other types in the population have a lesser ability; if others have a greater ability, the first type will be selected against, and if there are no fitness differences at all, there will be no natural selection at all.

If this explanation of how trait fitness can be a causal factor in the population is right, then Sober's argument is simply not applicable to the propensity interpretation of fitness, which is not overall fitness at all but rather the fitness of organisms with a certain trait, which Sober seems to grant is unproblematically causal[27] (again, recall that to claim that there is a propensity is just to claim that there is an underlying physical basis for the behavior). That is, the propensity of fitness would be causally efficacious, not causally inert.

Furthermore, even if it turned out that the overall fitness of an *organism* is causally inert, it might still be the case that overall fitness *differences* between organisms are causally efficacious. One particular organism may survive a predator rather than survive a disease (supposing that the organism was never exposed to the disease in question), but in the population as a whole, one would expect some organisms to survive predators, others to survive disease, etc. Thus, the overall fitness differences – more properly a property of the population rather than a property of any individual organisms – would be causally relevant to the changes in the distribution of organisms over time.[28] It still might turn out that in a particular generation, no organism was exposed to a given disease – but here it is important to keep in mind that since fitness is relative to the environment, an environment that lacks the relevant disease is one in which immunity does not contribute to overall fitness. And an environment with very little risk of disease is one in which immunity makes very little contribution to overall fitness.

[27] But see Sober (2013).

[28] This general point is argued for in more detail in Millstein (2006), but without referring to fitness directly; Sober (2013) seems to endorse it.

28.6 Conclusions

I have argued that the propensity interpretation of fitness, properly understood, not only solves the explanatory circularity problem and the mismatch problem, but also can withstand the Pandora's Box full of problems that have been thrown at it.[29] Fitness is the propensity (i.e., probabilistic ability, based on heritable physical traits) for organisms or types of organisms to survive and reproduce in particular environments and in particular populations for a specified number of generations; if greater than one generation, "reproduction" includes descendants of descendants. Fitness values can be described in terms of distributions of propensities to produce varying numbers of offspring and can be modeled for any number of generations using computer simulations, thus providing both predictive power and a means for comparing the fitness of different phenotypes. Fitness is a causal concept, most notably at the population level, where fitness differences are causally responsible for differences in reproductive success. Relative fitness is ultimately what matters for natural selection.

More generally, the above examination of fitness in the theory of natural selection implies the following lessons for understanding probability in other parts of evolutionary biology and other biological sciences.

First, if fitness and thus natural selection are probabilistic, then given the ubiquity of natural selection as an evolutionary process (not to mention random drift and other probabilistic evolutionary processes), we should *expect* to find probabilities throughout biology. Philosophers have only begun to scratch the surface in seeking out probabilities in the various areas of biology discussed at the outset of this paper. This means that there are potentially many discussions analogous to the one presented here; these should be explored. Furthermore, any failure to find the expected probabilities raises questions. If the study of an area of biology does not seem to invoke probabilities, why not? Does the aspect of organisms under study truly behave in a determinate way? Or are there other (pragmatic?) reasons why probabilities have not been invoked?

Second is a recommendation for the general approach that Mills and Beatty (1979) took, which has proved to be so fruitful. The stated definitions of biologists induced an apparent puzzle, the explanatory circularity problem. It is also often the case that the stated definitions of biologists can conflict. Thus, a promising way to understand concepts is to look at the way that biologists *use* their terms – that is, to look at actual biological practice. This holds for any sort of concept, but might be particularly useful for probabilistic ones, which can be difficult to express.

Third is not to lose sight of the original problems that motivate the understanding of a probabilistic concept. Technical issues (such as the mathematical ones discussed in this chapter) can be intriguing for philosophers and their examination is often productive, but they can also lead away from the core issues and cause one to misunderstand where the problem really lies. Philosophy of biology is most productive when it keeps both philosophy and actual biological practice in the forefront.

[29] See Drouet and Merlin (2014) for a recent argument that fitness should not be understood as propensities, and Pence and Ramsey (2013) for an alternative defense of it.

Finally, I hope to have shown that thinking about interpretations of probability in general, as well as the particular details of the invoked interpretation of probability, can be useful. Interpretations of probability provide criteria of things to look for and questions to ask. Figuring out which interpretations of probability are appropriate for different areas of biology is thus an essential first step. In the case of the propensity interpretation, it is important to understand that propensities are relational, that they are causal, and that they are always for a specified behavior; addressing these criteria leads to a better understanding of the nature of fitness. Other criteria can be identified for other interpretations of probability, such as whether degrees of belief (such as the uncertainties in medicine) are an inherent part of the invoked probabilities. In other words, interpretations of probability are more than just the answers to interesting philosophical questions; they can be useful tools.

Acknowledgments

Thanks to Marcel Weber for extremely helpful comments on this chapter, not to mention our many enjoyable discussions about the concept of fitness over the years. Thanks also to the Griesemer/Millstein Lab at UC Davis and John Jackson for insightful discussion and comments; to Jon Hodge for helpful comments and many conversations about fitness, chance, and natural selection; and of course to John Beatty for always gently pushing me to think harder about all things related to evolution and chance.

References

Abrams, Marshall (2006) Infinite Populations and Counterfactual Frequencies in Evolutionary Theory. *Studies in History and Philosophy of Biological and Biomedical Sciences*. 37. pp. 256–68.

Abrams, Marshall (2007) Fitness and Propensity's Annulment. *Biology and Philosophy*. 22. pp. 115–30.

Ariew, André and Ernst, Zachary (2009) What Fitness Can't Be. *Erkenntnis*. 71. pp. 289–301.

Ariew, André and Lewontin, Richard C. (2004) The Confusions of Fitness. *British Journal for the Philosophy of Science*. 55. pp. 347–63.

Beatty, John (1984) Chance and Natural Selection. *Philosophy of Science*. 51. pp. 183–211.

Beatty, John and Finsen, Susan (1989) Rethinking the Propensity Interpretation: A Peek Inside Pandora's Box. In Ruse, M. (ed.) *What the Philosophy of Biology Is*. pp. 17–30. Dordrecht: Kluwer.

Bouchard, Frédéric (2011) Darwinism Without Populations: A More Inclusive Understanding of the 'Survival of the Fittest'. *Studies in History and Philosophy of Biological and Biomedical Sciences*. 42. pp. 106–14.

Bouchard, Frédéric and Rosenberg, Alex (2004) Fitness, Probability, and the Principles of Natural Selection. *British Journal for the Philosophy of Science*. 55. pp. 693–712.

Brandon, Robert N. (1978) Adaptation and Evolutionary Theory. *Studies in the History and Philosophy of Science*. 9. pp. 181–206.

Brandon, Robert N. (1990) *Adaptation and Environment*. Princeton, NJ: Princeton University Press.

Brandon, Robert and Beatty, John (1984) The Propensity Interpretation of 'Fitness': No Interpretation Is No Substitute. *Philosophy of Science*. 51. 2. pp. 342–7.

Burian, Richard M. (1983) Adaptation. In Greene, M. (ed.) *Dimensions of Darwinism*. Cambridge: Cambridge University Press.

Colyvan, Mark (2005) Probability and Ecological Complexity. *Biology and Philosophy*. 20. pp. 869–79.

Craver, Carl (2007) *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. Oxford: Clarendon Press.

Darwin, Charles (1859) *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life.* 1st edition. London: John Murray.

Darwin, Charles (1871/1981) *The Descent of Man, and Selection in Relation to Sex*. Princeton, NJ: Princeton University Press.

Djulbegovic, Benjamin, Hozo, Iztok, and Greenland, Sander (2011) Uncertainty in Clinical Medicine. In Gifford, F. (ed.) *Philosophy of Medicine*. pp. 299–356. Oxford: Elsevier.

Drouet, Isabelle and Merlin, Francesca (2014) The Propensity Interpretation of Fitness and the Propensity Interpretation of Probability. *Erkenntnis*. pp. 1–12.

Fox Keller, Evelyn (1987) Reproduction and the Central Project of Evolutionary Theory. *Biology and Philosophy*. 2. pp. 383–96.

Gayon, Jean (1998) *Darwinism's Struggle for Survival: Heredity and the Hypothesis of Natural Selection*. Cambridge: Cambridge University Press.

Gillespie, John H. (1974) Natural Selection for Within-Generation Variance in Offspring Number. *Genetics*. 76. pp. 601–6.

Gillespie, John H. (1977) Natural Selection for Variances in Offspring Numbers–A New Evolutionary Principle. *American Naturalist*. 111. pp. 1010–14.

Gillies, Donald (2000) *Philosophical Theories of Probability*. London: Routledge.

Hájek, Alan (2012) Interpretations of Probability. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. Summer. [Online] Available from: http://plato.stanford.edu/entries/probability-interpret/. [Accessed 23 Sep 2015]

Hodge, M. J. S. (1987) Natural Selection as a Causal, Empirical, and Probabilistic Theory. In Krüger, L., Gigerenzer, G., and Morgan, M. S. *The Probabilistic Revolution*. pp. 233–70. Cambridge, MA: MIT Press.

Jackson, John P., Jr. (ms.) The Survival of the Unfit: Was Eugenics a Darwinian Discourse? Unpublished manuscript.

Kitcher, Philip (1987). Why Not the Best? In Dupré, J. (ed.) *The Latest on the Best: Essays on Evolution and Optimality*. pp. 77–102. Cambridge, MA: MIT Press.

Lennox, James G. and Wilson, Bradley E. (1994) Natural Selection and the Struggle for Existence. *Studies in the History and Philosophy of Science*. 25. 1. pp. 65–80.

Matthen, Mohan and Ariew, André (2002) Two Ways of Thinking about Fitness and Natural Selection. *The Journal of Philosophy*. 99. 2. pp. 55–83.

Mayr, Ernst (1988) *Toward A New Philosophy of Biology. Observations of an Evolutionist.* Cambridge, MA: Harvard University Press.

Merlin, Francesca (2010). Evolutionary Chance Mutation: A Defense of the Modern Synthesis' Consensus View. *Philosophy and Theory in Biology*. 2. e103. pp. 1–22.

Mills, Susan K. and Beatty, John H. (1979) The Propensity Interpretation of Fitness. *Philosophy of Science*. 46. 2. pp. 263–86.

Millstein, Roberta L. (2002) Are Random Drift and Natural Selection Conceptually Distinct? *Biology and Philosophy*. 17. 1. pp. 33–53.

Millstein, Roberta L. (2003a) Interpretations of Probability in Evolutionary Theory. *Philosophy of Science*. 70. 5. pp. 1317–28.

Millstein, Roberta L. (2003b) How Not to Argue for the Indeterminism of Evolution: A Look at Two Recent Attempts to Settle the Issue. In Hüttemann, A. (ed.) *Determinism in Physics and Biology*. pp. 91–107. Paderborn: Mentis.

Millstein, Roberta L. (2006) Natural Selection as a Population-Level Causal Process. *The British Journal for the Philosophy of Science*. 57. 4. pp. 627–53.

Millstein, Roberta L. (2009) Populations as Individuals. *Biological Theory*. 4. 3. pp. 267–73.

Millstein, Roberta L. (2010) The Concepts of Population and Metapopulation in Evolutionary Biology and Ecology. In Bell, M. A., Futuyma, D. J., Eanes, W. F., and Levinton, J. S. (eds.) *Evolution Since Darwin: The First 150 Years*. pp. 61–86. Sunderland, MA: Sinauer.

Millstein, Roberta L. (2011) Chances and Causes in Evolutionary Biology: How Many Chances Become One Chance. In Illari, P. McKay, Russo, F., and Williamson, J. (eds.) *Causality in the Sciences*. pp. 425–44. Oxford: Oxford University Press.

Millstein, Roberta L. (2012) Darwin's Explanation of Races by Means of Sexual Selection. *Studies in History and Philosophy of Biological and Biomedical Sciences*. 43. pp. 627–33.

Millstein, Roberta L. and Skipper, Robert A. (2007) Population Genetics. In Hull, D. and Ruse, M. (eds.) *The Cambridge Companion to the Philosophy of Biology*. pp. 22–43. Cambridge: Cambridge University Press.

Pence, Charles H. and Ramsey, Grant (2013) A New Foundation for the Propensity Interpretation of Fitness. *British Journal for the Philosophy of Science*. 64. pp. 851–81.

Pigliucci, Massimo and Kaplan, Jonathan (2006) *Making Sense of Evolution: The Conceptual Foundations of Evolutionary Biology*. Chicago, IL: University of Chicago Press.

Popper, Karl R. (1959) The Propensity Interpretation of Probability. *The British Journal for the Philosophy of Science*. 10. 37. pp. 25–42.

Ramsey, Grant (2013) Can Fitness Differences Be a Cause of Evolution? *Philosophy and Theory in Biology*. 5. e401.

Rice, Sean (2008) A Stochastic Version of the Price Equation Reveals the Interplay of Deterministic and Stochastic Processes in Evolution. *BMC Evolutionary Biology*. 8. p. 262.

Rice, Sean and Papadopoulos, Anthony (2009) Evolution with Stochastic Fitness and Stochastic Migration. *PLoS ONE*. 4. 10. pp. 1–12.

Rice, Sean, Papadopoulos, Anthony, and Harting, John (2011) Stochastic Processes Driving Directional Evolution. In Pontarotti, P. (ed.) *Evolutionary Biology: Concepts, Methods, Macroevolution*. pp. 21–33. Berlin: Springer-Verlag.

Richardson, Robert C. and Burian, Richard M. (1992) A Defense of Propensity Interpretations of Fitness. *PSA 1992: Proceedings of the Biennial Meeting of the Philosophy of Science Association*. 1. pp. 349–62.

Rosenberg, Alex (1985) *The Structure of Biological Science*. Cambridge: Cambridge University Press.

Rosenberg, Alex and Williams, Mary B. (1986) Discussion: Fitness as a Primitive and Propensity. *Philosophy of Science*. 53. pp. 412–18.

Scriven, M. (1959) Explanation and Prediction in Evolutionary Theory. *Science*. 130. pp. 477–82.

Sober, Elliott (1984) *The Nature of Selection*. Chicago, IL: University of Chicago Press.

Sober, Elliott (2001) The Two Faces of Fitness. In Singh, R. S., Krimbas, C. B., Paul, D. B., and Beatty, J. (eds.) *Thinking About Evolution: Historical, Philosophical, and Politcal Perspectives*. pp. 309–21. Cambridge: Cambridge University Press.

Sober, Elliott (2013) Trait Fitness Is Not a Propensity, but Fitness Variation Is. *Studies in History and Philosophy of Biological and Biomedical Sciences*. 44. pp. 336–41.

Walsh, Denis M. (2007) The Pomp of Superfluous Causes: The Interpretation of Evolutionary Theory. *Philosophy of Science*. 74. pp. 281–303.

Waters, Kenneth C. (1986) Natural Selection without Survival of the Fittest. *Biology and Philosophy*. 1. pp. 207–25.

PART VII

APPLICATIONS OF PROBABILITY: PHILOSOPHY

CHAPTER 29

PROBABILITY IN EPISTEMOLOGY

MATTHEW KOTZEN

29.1 Introduction

Increasingly in recent years, formal approaches to epistemology have become prevalent, especially ones that involve some sort of appeal to probabilities. Typically, advocates of these approaches do not claim that formal approaches could or should entirely replace traditional epistemological approaches; rather, they usually claim only that formal tools can be used to supplement traditional techniques in order to precisify various epistemological problems, theses, and arguments. In this chapter, I survey some of the areas of epistemology in which appeals to probability have been most influential.

29.2 Full and Partial Beliefs

The "traditional" concept of a belief is a binary concept; for all A, one either believes A or fails to believe A. Much of traditional epistemology, then, focuses on the question of what it takes for a belief that A to be *justified*, or to be an item of *knowledge*, etc. Another way of conceiving of belief is as a *graded* notion; on this conception, one's doxastic attitude toward A can be represented by a real number in the $[0,1]$ interval—referred to as one's "credence" in A—where higher credences correspond to "more" belief in A. There are different stories that we might tell about how these conceptions of belief are related; I will discuss four such views.

On one view, there are psychological states that correspond to binary beliefs, and there are different psychological states that correspond to credences; thus, the two conceptions are not competitors for the correct account of the metaphysics of doxastic attitudes, but are rather accounts of two different sorts of doxastic attitudes. But there is clearly *some* metaphysical connection between these attitudes; it would be hard to make sense of an agent who claimed to have credence 0 in a proposition A and yet claimed to believe that A. Moreover, some philosophers have worried that this view simply isn't psychologically

realistic, since "[t]here is no evidence to believe that the mind contains two representational systems, one to represent things as being probable or improbable and the other to represent things as being true or false."[1]

On a second view of the relation between binary beliefs and credences, credences are to be analyzed as full beliefs in explicit claims about objective probabilities; thus, for example, a credence of .6 in the proposition that it is going to rain tomorrow is to be analyzed as a full belief in the proposition that the objective probability of rain tomorrow is 60%.[2] But it's not clear that there is a coherent notion of objective probability that can support this analysis in all cases; for example, I might have a credence of .7 that Tom was at the party yesterday, but it's at least somewhat natural to think that the objective probability that Tom was at the party yesterday is either 1 or 0, depending on whether he was there or not. Similarly, I might have a credence of .7 in the proposition that a particle will decay in the next hour because I'm 50% sure that the particle has an objective probability of 80% of decaying, and 50% sure that the particle has an objective probability of 60% of decaying; in such a case, though my credence that the particle will decay is .7, I definitely do *not* believe that the objective probability that the particle will decay is 70%. Again, it's not clear that the view under consideration can be extended to this sort of case. Someone might respond to both of these cases by pointing out that there is a coherent *evidential* notion of objective probability on which the objective probability of Tom's being at the party yesterday is .7, and on which the objective probability of the particle decaying is .7.[3] But this doesn't seem to lead to a plausible theory of the relation between binary beliefs and credences either; my credence that Tom was at the party yesterday still might be .7, even though I don't have *any* beliefs at all about evidential probabilities.

On a third view, binary belief is to be analyzed in terms of credence; on the most natural way of developing this view, to have a binary belief that A is to have a credence in A that is above a certain threshold, where that threshold is determined at least in part by either the context of the believer or the context of the belief-ascriber. Kaplan 1996 refers to this view as the "threshold view." Two significant problems arise for this view, both of which are discussed in Stalnaker 1984. First, it is hard to see which features of either the believer's or the belief-ascriber's context could make it the case that the threshold value for belief is, say, .9786 rather than .9785. Some philosophers (such as Foley 1993 chapter 4 and Hunter 1996) have argued that the threshold must therefore be vague, but it is not completely clear whether this helps.[4] Second, if an agent regards A and B to be independent, then it is possible for her to have a rational credence in A above a given threshold, and also to have a rational credence in B above that threshold, and yet to have a rational credence in $A \wedge B$ that is below the threshold. If binary belief is just credence above the relevant threshold, then this situation corresponds to one in which the agent believes A, and also believes B, but fails to believe $A \wedge B$. This is somewhat odd, since we usually think of an agent who believes A and B separately but who fails to believe $A \wedge B$ as exhibiting some rational failing. This issue has received much attention in discussions of the closely-related Preface Paradox—see, e.g., Christensen 2004—and the Lottery Paradox—see, e.g., Kyburg 1961.

[1] Weatherson 2005, p. 420.

[2] See, e.g., Frankish 2009, p. 77.

[3] This notion of "evidential probability" is discussed in Section 29.6.

[4] See, e.g., Weatherson 2005, pp. 420–1.

On a fourth view, credences are understood to *replace* binary beliefs. On this view, there just are no psychological states corresponding to binary beliefs; rather, all that we have are credences in various propositions—some higher and some lower—but none of those credences stand out for special classification as our "beliefs."[5] Of course, it's often convenient to talk in terms of full belief, as when we say "I believe it's going to rain tomorrow," but strictly speaking, there are no psychological states that correspond to the binary notion. This view naturally raises the question of how to reconceive traditional epistemological concepts such as *knowledge* in the context of credences. Moss [2013] develops an answer to this question.

29.3 Synchronic Constraints

So far, the only constraint on credence functions that we have seen is the constraint that they are real numbers in the [0,1] interval. There is a variety of different further constraints that have been argued for. Some of these constraints are *synchronic*; they are constraints that are alleged to apply to agents *at a particular time*. Other constraints are *diachronic*; they apply to agents' *revisions* of their credences from one time to another. This section holds a discussion several proposed synchronic constraints on credences, and Section 29.4 discussion of several proposed diachronic constraints.

Even within the synchronic constraints on credences, we can distinguish constraints on our *unconditional* credences from constraints on our *conditional* credences. Unconditional credences are credences that we assign to individual propositions. Conditional credences, by contrast, are often expressed as credences in A, *given* B. Traditionally, a conditional credence $p(A|B)$ was understood to be defined as $\frac{p(A \wedge B)}{p(B)}$. More recently, Alan Hájek has argued that we should instead understand conditional credences as primitive; among other things, this understanding has the advantage of allowing conditional credences to be defined when $p(B) = 0$ (which the traditional definition clearly does not).[6] But the issue of how conditional credences are best to be understood is in principle separable from the question of how they are rationally constrained (even if these questions turn out to be related in various ways). Sections 29.3.1 and 29.3.2 develop some constraints on unconditional credences, and Sections 29.3.3 and 29.3.4 develop some constraints on conditional credences.

29.3.1 Probabilism

One of the most widely discussed synchronic constraints on unconditional credence is the constraint that credences obey the standard Kolmogorov probability axioms.[7] This constraint can be developed as follows: we start with a set of mutually exclusive and jointly exhaustive possibilities, which is designated Ω. Then, we define a field F of subsets of Ω; call

[5] Jeffrey 1970 (pp. 171–2) expresses sympathy for this view: "I am inclined to think Ramsey sucked the marrow out of the ordinary notion [of belief], and used it to nourish a more adequate view."

[6] Hájek 2003.

[7] Kolmogorov 1956.

the members of F "propositions." Finally, we define a probability function p from F to the [0,1] interval of the real numbers that obeys the following three axioms for all propositions A and B in F:

1. $p(A) \geq 0$.
2. $p(\Omega) = 1$.
3. If $A \cap B = \emptyset$, then $p(A \cup B) = p(A) + p(B)$.

The distinctive claim of Probabilism is that it is a rational constraint on credences that they be probability functions, in the sense defined above. So, according to Probabilism, a rational agent must (1) assign a non-negative credence to every proposition, (2) assign a credence of 1 to the set of all possibilities, and (3) where A and B are mutually exclusive, assign a credence to $A \vee B$ that is the sum of the credences she assigns to A and to B separately. A credence function obeying these constraints is called "coherent," and one not obeying these constraints is called "incoherent."[8]

A variety of different arguments have been offered for Probabilism. First, Synchronic Dutch Book Arguments proceed from the premise that agents with incoherent credences are committed in some way (in virtue of the Dutch Book Theorem) to a series of bets that jointly guarantee them a loss; but since (the Argument continues) being so committed is irrational, it must be a rational constraint on credences that they be probabilities.[9] Early Synchronic Dutch Book Arguments assumed that the nature of this "commitment" was one of identity; according to these arguments, to have a credence in A of .5 *just is* to be disposed to accept a bet at 1 : 1 odds that wins if and only if A is true. On this understanding, the Dutch Book Theorem shows that agents with incoherent credences are disposed to accept each in a series of bets such that the bets jointly guarantee that the agent will lose money. This version of the Synchronic Dutch Book Argument faced the objection that a rational agent might well have a credence in A of .5, say, and yet not be disposed to bet on A at 1 : 1 odds owing to a distaste for gambling, or owing to risk-aversion, etc. Thus, more recently, there have been a variety of "depragmatized" Dutch Book Arguments which claim only that an agent with a credence in A of .5 is *committed to the fairness of* a bet at 1 : 1 odds, regardless of whether he is actually disposed to accept such a bet.[10] According to these arguments, it is a rational failing to be committed to the fairness of a series of bets that jointly guarantee a loss; hence, since incoherent credences commit an agent in precisely this way, rational agents are constrained to have coherent credences.[11]

A second type of argument for Probabilism appeals to a "representation theorem."[12] Such theorems establish that, if an agent's preferences satisfy various plausible constraints such as transitivity and connectedness, then that agent is representable as having a utility function $U(x)$ and a coherent credence function $C(x)$ such that the agent prefers A to

[8] For a general discussion of this topic, see Lyle Zynda's chapter "Subjectivism" (17) in this volume (Zynda 2016).

[9] The original Synchronic Dutch Book Argument is usually attributed to Ramsey 1926 and de Finetti 1937.

[10] See Armendt 1993, Christensen 1996, and Skyrms 1984.

[11] See Easwaran 2011 for an excellent overview of Dutch Book Arguments.

[12] See Savage 1954, Jeffrey 1965, Joyce 1999, and Maher 1993.

B just in case her "expected utility" of *A*, calculated via $U(x)$ and $C(x)$, is higher than her expected utility *B*. Arguments for Probabilism based on representation theorems have faced a variety of objections. Many have focused on the worry that just because a rational agent's preferences are *representable* as being derived from a utility function $U(x)$ and a coherent credence function $C(x)$, it doesn't obviously follow that he *actually has* $C(x)$ as his credence function; in other words, even if the representation theorem argument establishes that all rational agents are representable as having coherent credence functions, that conclusion falls short of establishing the Probabilist claim that all rational agents *actually do have* coherent credence functions.[13] In particular, Zynda argues that alternative representations of an agent who satisfies the intuitive constraints—ones which *don't* appeal to a coherent credence function—are possible, so the significance of the fact that such agents can *also* be represented as having coherent credence functions isn't clear. Other objections have included criticism of the rational constraints on preferences (such as transitivity and connectedness) that give rise to the representation theorem.[14]

A third argument for Probabilism appeals to the notion of *calibration*.[15] To have credences that are perfectly calibrated is to have credences that perfectly match relative frequencies; proportion *x* of the cases in which you assign a credence of *x* to a proposition are cases in which that proposition is true, for all *x*. (So, 20% of the cases in which you assign a credence of .2 to a proposition are cases in which the relevant proposition is true, 30% of the cases in which you assign a credence of .3 to a proposition are cases in which the relevant proposition is true, etc.) However, even if you are not *perfectly* calibrated, it still makes sense to talk about being *more* or *less* calibrated; for instance, other things equal, if *A* is true in 21% of the cases in which I assign a credence of .2 to *A*, my credence function is better calibrated than it would be if *A* were true in 22% of the cases in which I assign a credence of .2 to *A*. The argument for Probabilism, then, shows that if your credence function is incoherent, then there is some coherent credence function that is better calibrated than yours under any logically consistent assignment of truth-values to propositions; from this fact, it follows, allegedly that you're rationally constrained to have a coherent credence function.[16] However, some philosophers have expressed doubt whether perfect calibration is really a rational ideal.[17] Moreover, one might think that what really matters is how well-calibrated your credence function is in the *actual* world, not how well-calibrated it is in *all possible worlds*, and coherence does not guarantee maximal calibration in the actual world.[18]

Finally, a fourth argument for Probabilism, due to Joyce 1998, alleges that for any incoherent credence function, there corresponds a coherent credence function that is strictly more accurate under every logically consistent assignment of truth-values to propositions. This argument is similar in some ways to the calibration argument above, but whereas the calibration argument appeals to relative frequencies, Joyce's argument appeals to the *accuracy* of credences. Non-extreme credences, of course, can't be accurate in the same way that binary beliefs can be (since a binary belief is accurate iff it is true, and there's no

[13] See, e.g., Hájek 2008.
[14] See, e.g., Schick 2003.
[15] See van Fraassen 1983 and Shimony 1988 for examples of this argument strategy.
[16] See Joyce 2004 for an explanation of this style of argument.
[17] See Seidenfeld 1985, Joyce 1998, and Hájek unpublished A.
[18] Hájek 2008.

obvious sense in which a non-extreme credence can be true); however, as in the calibration argument, we can say that when A is true, a credence in A is more accurate the closer it is to 1. Joyce's argument then proceeds to show that incoherent credence functions are always "less accurate than they could be," since whenever an agent's credence function is incoherent, he has available to him some coherent credence function which is more accurate than his credence function, no matter how the world turns out to be. Important objections to Joyce's argument are raised in Bronfman unpublished, Hájek 2008, and Maher 2002; some of these objections parallel the objection discussed above to the calibration argument by arguing that what really matters is how accurate your credence function is in the *actual* world, rather than *all possible worlds*. Joyce responds to these and other objections in Joyce 2009.

29.3.2 Regularity

Another proposed rational constraint on unconditional credence is the constraint that, if an agent regards a proposition A to be possible, then she should assign positive credence to A. The nature of the possibility at issue here is controversial; some authors have logical possibility in mind, whereas others seem to appeal to metaphysical possibility.[19]

But however we understand it, this constraint isn't entailed by the probability axioms; all that axiom 1 from Section 29.3.1 requires is that the agent assign *non-negative* credence to each proposition. And it's at least very plausible that an agent is permitted (perhaps even *required*) to assign a credence of 0 to propositions that she regards to be *impossible*: the proposition that she doesn't exist, for instance, or the proposition that 1+1=17. The constraint under consideration here entails that propositions that an agent regards to be impossible are the *only* propositions to which she is rationally permitted to assign a credence of 0. That constraint is called Regularity.[20]

On the one hand, Regularity can seem plausible. It's natural to think that your credence function should distinguish between those propositions that you regard as possible and those that you regard as impossible; however, if there are propositions of each type to which you assign credence 0, then it looks as if your credence function isn't able to make this distinction in all cases. Similarly, if I believe that B is true in only a proper subset of the worlds in which A is true, then it is natural to think I ought to assign a higher credence to A than to B; after all, I think that there are ways for A to be true while B is false, but no ways for B to be true while A is false. But this natural thought looks to entail Regularity. If, for any A and B, my believing that there are possibilities compatible with A but not with B (and not vice versa) rationally requires my assigning a higher credence to A than to B, then that must be because I am required to assign positive credence to the (non-empty) set of possibilities in which A is true but B is false, but I am required to do that only if I am required to assign positive credence to every proposition that I regard as possible.

[19] The versions of Regularity discussed in Shimony 1970 and Skyrms 1980 appeal to the former understanding of possibility, whereas Lewis 1980 (at least arguably) appeals to the latter. For an overview of the options, see Hájek 2012 and Hájek unpublished B.

[20] Versions of the Regularity constraint have been proposed by Kemeny 1955, Shimony 1955, Shimony 1970, Jeffreys 1961, Edwards et al. 1963, Carnap 1963a, Stalnaker 1970, Lewis 1980, Skyrms 1980, Appiah 1985, Jackson 1987, and Jeffrey 1992.

Regularity runs into trouble, however, in certain cases with uncountably many possible outcomes. Suppose that I am about to throw an infinitely sharp dart at a dartboard, and suppose for simplicity that the dart is guaranteed to hit a spot somewhere on the dartboard. Now, consider the uncountable set of all of the points on the dartboard; for each such point, I regard the proposition that the dart lands on precisely that point to be possible. But there is no way for a coherent credence function to assign a (real-valued) positive credence to each of these propositions; indeed, as Hájek 2003 shows, any probability function defined on an uncountable algebra assigns probability 0 to uncountably many propositions. A similar problem arises for unmeasurable sets of points. Consider some unmeasurable set of points on the dartboard *S*, and consider the proposition *B* that the dart lands on one of the points in *S*. As before, I regard *B* to be possible; after all, the dart could certainly hit any individual point in *S*, in which case *B* would be true. But again, there is no way for a coherent credence function that respects the symmetries of the situation to assign a (real-valued) positive credence to *B*.[21] In response to these complications, some have proposed a relaxation of the requirement that credence functions be real-valued, suggesting instead that we should allow hyperreal-valued credence functions.[22] Williamson 2007 and Hájek unpublished B raise various objections to this proposal.

29.3.3 Reflection

Another alleged synchronic constraint on rationality is van Fraassen's "Reflection Principle."[23] Here is van Fraassen's statement of the "General Reflection Principle": "My current opinion about event E must lie in the range spanned by the possible opinions I may come to have about E at later time *t*, as far as my present opinion is concerned."[24] This constraint is naturally understood as a constraint on an agent's conditional credences: her current credence in *A*, conditional on the proposition that her future credence in *A* will be in the [m,n] range, should itself be somewhere in the [m,n] range.

This principle is motivated by the thought that you are rationally obligated to treat your future self as a better authority with regard to the subject matter of your beliefs than you are. So, conditional on the information that your future self is going to have a credence in *A* that is in some range, it can seem odd to assign a credence to *A* that is outside that range rather than to "respect" your future self's epistemic authority by assigning a conditional probability that conforms to hers. As a result, the thought goes, even if you aren't actually certain of what your future credences will be, you are still constrained to have *conditional* credences in each proposition that obey General Reflection.

Of course, General Reflection is much less plausible in cases where an agent rationally believes that at some time between now and the time at which her future self has a credence in *A* that is in the [m,n] range, she is going to lose some of her current evidence, or else

[21] See Hájek unpublished B for a discussion of complications arising from unmeasurable sets and sets of measure 0.

[22] See, e.g., Lewis 1980, who cites Bernstein and Wattenberg 1969 for their technical construction involving infinitesimals.

[23] See van Fraassen 1984 and van Fraassen 1995.

[24] van Fraassen 1995, p. 16.

become irrational or suffer some other sort of epistemic defect; for instance, I might have a rational conditional credence in A, conditional on my getting kidnapped tonight and brainwashed into having a credence in A of .8 by tomorrow, of .5.[25] But in cases where an agent rationally believes that she won't have lost any evidence that she currently has or suffer any other sort of epistemic deficit by t, it is more plausible that she is rationally constrained to obey General Reflection. This constraint clearly isn't entailed by the agent having a probabilistic credence function, so van Fraassen thinks that Reflection constitutes a substantial additional synchronic constraint on rationality.

Various objections have been raised against Reflection. Elga 2000 argues for the "Thirder" answer to the Sleeping Beauty Problem (discussed in Section 29.4.2 below), and claims that that answer is incompatible with Reflection. Arntzenius 2003 discusses several cases where it seems as though a rational agent violates Reflection even though no *actual* epistemic deficit befalls the agent; in Artnzenius's cases, it seems to be enough that the agent's future self will have some positive credence that the relevant deficit has been suffered (and that the agent's previous self knows this), even if no deficit is ever actually suffered by the agent.

29.3.4 The Principal Principle

In Lewis 1980, David Lewis defended a synchronic rational constraint on conditional credences that, he argues, is imposed by information about objective chances.[26] The guiding intuition behind this constraint is that a rational agent's conditional credence at t that A is true, conditional on the information that A's objective chance at t is x, should itself be equal to x; for instance, a rational agent's conditional credence at 1:00 that the coin will land heads at 2:00, conditional on the information that the chance at 1:00 that the coin will land heads at 2:00 is .423, should itself be .423.

However, Lewis claims, a rational agent's conditional credence at t that A is true, conditional on the information that A's objective chance at t is x *and* that E is true, might not be x; in cases where E contains "inadmissible" information, the agent's conditional credence might take a value other than x. For Lewis, admissible information is "the sort of information whose impact on credence comes entirely by way of credence about the chances of those outcomes"[27]; a prime example of inadmissible information is information about the future. For instance, suppose that E is the information that a known-to-be-reliable crystal ball predicts, at 1:00, that the coin will land heads at 2:00. Then, the Principal Principle doesn't entail that a rational agent's conditional credence at 1:00 that the coin will land heads at 2:00, conditional on the information that the objective chance at 1:00 that the coin will land heads at 2:00 is .5 *and* that E is true, is .5; in this case, the rational agent's conditional credence might be 1, for instance.

[25] See Christensen 1991 and Talbott 1991 for discussion of cases like these.

[26] Lewis 1980 contains the canonical formulation of this type of constraint, though Lewis there acknowledges his debt to Mellor 1971, which (in Lewis's words) "presents a view very close to" Lewis's own view.

[27] Lewis 1980, p. 92.

Following Lewis 1980: Let p be any "reasonable initial credence function."[28] Let x be any real number in the [0,1] interval. Let X be the proposition that the objective chance at t of A's holding is x. And let E be any proposition that is both compatible with X and admissible at t. Then, the Principal Principle says that p will satisfy the constraint that $p(A|X \wedge E) = x$. In other words: as long as an agent's initial credence function is reasonable, her initial conditional credence in A, conditional on the conjunction of a proposition specifying the objective chance of A at t with any (compatible) admissible proposition, is just the value that was specified to be the objective chance of A at t.

Later, Lewis worried that the Principal Principle conflicted with his own Humean "best system" theory of laws and chances, owing to the so-called "problem of undermining futures."[29] The source of this problem is that, if a Humean theory of chance—according to which present chances supervene on the whole of history, including the future as well as the present and the past—is true, then at any time t, there are non-actual futures with non-zero chances which are such that, were they to transpire, some of the chances at t would have different values from the ones they actually have. For instance, suppose that a coin is flipped a large but finite number of times in the entire history of the universe, and consider some time t, by when only some small fraction of the flips have taken place. Suppose that the coin actually lands heads (approximately) half the time in the entire history of the universe, so that the best theory of chance entails that the chance of any particular coin flip landing heads is .5. Still, according to that theory, the "undermining future" F in which the coin land heads on each flip after t has some positive chance; and if such a future were to obtain, then the best theory of chance would entail that the chance at t of any particular coin flip after t landing heads is greater than .5 (and hence different from its actual value of .5).

So far, Lewis thinks that this consequence is merely "peculiar," but the conflict with the Principal Principle arises when we consider the chance at t of F itself. The actual chance at t of any particular flip after t landing heads is, we're supposing, .5. Suppose that there are n flips which occur after t. Then, the actual chance at t of F is $.5^n$. But if F were to occur, then the chance at t of any particular flip after t landing heads would be greater than .5, and hence the chance of t of all n flips after t landing heads would be m^n, for some $m > .5$. Let X be the proposition specifying the actual chance of F at t (i.e., $.5^n$), and let E be any admissible proposition compatible with X. By the Principal Principle, a rational agent's $p(F|X \wedge E)=.5^n$, which is positive (since n is finite). But, since F entails a different chance at t of F than X entails, F is inconsistent with X, and hence a rational agent's $p(F|X \wedge E) = 0$. Contradiction.

One possible response would be to say that information about objective chances (such as X in the previous paragraph) is itself information about the future, and hence inadmissible.

[28] Lewis requires that a reasonable initial credence function at least obey the axioms of the probability calculus (see Section 29.3.1 above) and that it be regular (see Section 29.3.2 above). He also says that a reasonable initial credence function is such that "if you always learned from experience by conditionalizing on your total evidence, then no matter what course of experience you might undergo your beliefs would be reasonable for one who had undergone that course of experience." (Lewis 1980, p. 268.) See Section 29.4.1 below for a discussion of conditionalization.

[29] Lewis develops this worry in Lewis 1986. Halpin 1994, Hall 1994, and Thau 1994 all advocate versions of the strategy of rejecting the Principal Principle in response to the problem. See Wolfgang Schwarz's chapter "Best Systems Approaches to Chance" (20) in this volume (Schwarz 2016), for further discussion.

But, as Lewis points out,[30] this would render the Principal Principle useless; the whole point of the Principal Principle is to articulate the rational constraint imposed by information about current objective chances, and if Lewis's Humean theory of chances is correct, then current chances supervene in part on future events.

In Lewis 1994, Lewis qualifies the Principal Principle and endorses a modified "New Principle." Recall that the original Principal Principle says that, where X is the proposition that the objective chance at t of A's holding is x, a rational agent's $p(A|X \wedge E) = x$. The New Principle modifies the content of X; on the modified understanding, X is the proposition that the objective chance at t of A's holding, *conditional on the actual true theory of chance*, is x. We therefore ignore undermining futures which would contradict the actual true theory of chance.[31] However, some authors have argued that Lewis didn't actually need to abandon the original Principal Principle in favor of the New Principle.[32]

29.4 Diachronic Constraints

29.4.1 Conditionalization

Perhaps the most widely discussed diachronic constraint on credences is the Rule of Conditionalization. According to this rule, when you learn E and nothing more, you should update your credence in any proposition A from your old unconditional credence $p(A)$ to your old conditional credence $p(A|E)$. It is somewhat tempting to read this constraint as a trivial one, since it is tempting to understand $p(A|E)$ as "the credence that it's rational to assign to A, given that you've learned that E is true." But this is a mistake; $p(A|E)$ is an entirely synchronic feature of your old credence function, and doesn't by itself entail anything about how you should *update* your credence function when you acquire new evidence such as E. On the "traditional" view of conditional credences discussed in Section 29.3, $p(A|E)$ is just an abbreviation for $\frac{p(A \wedge E)}{p(E)}$; and it is clearly a substantive constraint on rationality that, when a rational agent learns that E, her new credence in A should be the ratio of her old credence in $A \wedge E$ to her old credence in E. And even if Hájek is right that conditional credences should be regarded as primitive rather than defined in the manner above, it is *still* a substantive constraint on rationality that, when you learn E, your *new* credence in A should be equal to your *old* primitive conditional credence $p(A|E)$.

One prominent argument that is offered in support of the Rule of Conditionalization is the Diachronic Dutch Book Argument, which—though similar in strategy—is importantly different from the Synchronic Dutch Book Argument discussed in Section 29.3.1. The first Diachronic Dutch Book Argument appeared in Teller 1973, though Teller attributed the argument to David Lewis.[33] Whereas the Synchronic Dutch Book Argument assumes that an agent who has a $p(A) = x$ is committed to regarding as fair a bet on A at $x : 1 - x$ odds,

[30] Lewis 1994, pp. 485-6.

[31] See Loewer 2004 and Roberts 2001 for clear expositions of the New Principle.

[32] See, e.g., Vranas 1998 and Roberts 2001.

[33] Lewis later developed a version of the Diachronic Dutch Book Argument in Lewis 1999.

the Diachronic Dutch Book Argument makes the additional assumption that an agent with $p(A|B) = x$ is committed to regarding as fair a *conditional* bet which is "called off" if B is false, and which becomes an ordinary bet on A at $x : 1 - x$ odds if B is true. Then, the argument proceeds by showing that an agent who violates the Rule of Conditionalization is committed to regarding as fair a series of bets, some made before learning B and some made after learning B, which jointly guarantee a net loss for the agent over time. In general, philosophers have found Diachronic Dutch Book Arguments to be less persuasive than their Synchronic cousins.[34]

Another important argument for the Rule of Conditionalization is developed in Greaves and Wallace 2006. The strategy of this argument is to define a notion of the "epistemic utility" of a credence according to which, on the assumption that A is true, a credence of x in A has higher utility than a credence of y in A just in case $x > y$. (On the assumption that A is false, a credence of x in A has more epistemic utility than a credence of y in A just in case $y > x$.) Then, Greaves and Wallace define an agent's *expected* epistemic utility, which depends on both the agent's credence distribution and the epistemic utilities of those credence functions. Greaves and Wallace then prove that under certain conditions, the Rule of Conditionalization is the unique update rule that maximizes expected epistemic utility.

There are some approaches to epistemic rationality that reject the need for any update rule at all. For example, in Williamson 2010, Jon Williamson defends a version of "Objective Bayesianism" that features three synchronic constraints on a rational agent's credences. But instead of endorsing any update rule that constrains an agent to *change* his credences in any particular way, Williamson argues that the three synchronic constraints should just be applied to each new evidential situation that the agent finds herself in. Thus, different evidential situations constrain an agent to have different credences, but not because new evidence mandates a *change* in credences *as such*; rather, new evidence creates a new evidential situation in which the three synchronic norms generate new synchronic constraints.

29.4.2 Updating on *De Se* Information

On the traditional understanding, credences are interpreted as being assigned to *uncentered* propositions, or propositions about what the world is like that reflect a "third-personal" perspective; we then address both synchronic questions about how an agent should assign his credences to various uncentered propositions, and also diachronic questions about how he should update his credences in those uncentered propositions as he collects new evidence. But in recent years, there has been considerable interest in the question of how an agent should apportion and update the credences he assigns to *centered* propositions, or propositions which reflect a "first-personal" perspective by being about the location or identity of the believer of the proposition.[35] It's plausible that to know which uncentered propositions are true isn't always to know everything about your situation, since you might

[34] See Christensen 1991 for one line of resistance to Diachronic Dutch Book Arguments.

[35] The seminal articles on centered possible worlds are Quine 1969, Perry 1979, Lewis 1979, and Stalnaker 1981. The rebirth of interest in the topic began with Elga 2000 and Lewis 2001. For a general discussion, see Mike Titelbaum's chapter "Self-Locating Credences" (31) in this volume (Titelbaum 2016).

not know *who* or *where* or *when* you are in the possible world in which those uncentered propositions are true. For instance, you might know that the actual world is one with two beings in it, one who lives on the tallest mountain and one who lives on the coldest mountain, and yet not know which being *you* are.[36] And yet it seems that you might have rational credences in each of these "centered" possibilities, so the question arises of what the rational constraints on these credences are.

This question, and various answers to it, have been precisified in discussions of the "Sleeping Beauty Problem."[37] In this case, Beauty is told on Sunday that she is about to be put to sleep, and that a fair coin will be flipped (which she will not see). If the coin lands heads, she will be awakened only on Monday. If the coin lands tails, she will be awakened on Monday, then have her memory of the Monday waking erased, then be put back to sleep, and then be awakened again on Tuesday. Later, she is awakened. Question: When she is awakened, what should her credence be that the coin landed heads? Some philosophers ("Halfers") have defended the answer "one-half," whereas others ("Thirders") have defended the answer "one-third." When Beauty awakes, there are now two centerings of the tails-world that Beauty regards to be possible ("The coin landed tails and it's Monday" and "The coin landed tails and it's Tuesday"), and only one centering of the heads-world ("The coin landed heads and it's Monday") that she regards to be possible. According to the Halfer position, this is no reason for Beauty to increase her credence in the tails-world from $\frac{1}{2}$ to $\frac{2}{3}$ (and hence no reason to reduce her credence in the heads-world from $\frac{1}{2}$ to $\frac{1}{3}$). By contrast, the "Thirder" position entails that Beauty *should* increase her credence in the tails-world in response to the fact that she is now in a situation where there are two centerings that are compatible with the tails-world (and where there is still only one centering that is compatible with the heads-world).

29.5 The Requirement of Total Evidence

The Requirement of Total Evidence (RTE) is plausibly interpreted to have both a synchronic and a diachronic component, so it doesn't fit completely naturally into either of the previous two sections. The RTE enjoins us to always consider *all* of our evidence, rather than just some incomplete part of it. The synchronic constraint generated by the RTE is the constraint that the credences that a rational agent has any any time t should be the credences that are justified by his total evidence at time t. Suppose, for instance, that it's rational for you to believe at t that any particular healthy person is unlikely to die in the next year, but that healthy people *who often skydive* are likely to die in the next year. If you know that John is a healthy skydiver, then your credence that John is going to die in the next year should have the higher value justified by your total evidence which includes the fact that John skydives, rather then the lower value that would be justified by the less-than-total evidence which includes the fact that John is healthy but excludes the fact that John is a skydiver.

[36] This example comes from Lewis 1979.

[37] See, e.g., Elga 2000, Lewis 2001, Dorr 2002, Arntzenius 2002, Hitchcock 2004, and Meacham 2008.

The diachronic constraint imposed by the RTE is the constraint that your *revisions* of your credences should be the ones that are justified by a consideration of your *total* new evidence, rather than some partial component of it. This constraint is often discussed in the context of the Rule of Conditionalization; the idea is that, when you acquire some new evidence, your new credence in each A should be your old conditional credence $p(A|E)$, where E represents a *total* statement of the new evidence that you've acquired. So, for instance, when you know nothing about John and then learn that John is a healthy skydiver (and nothing more), your new credence that John will die in the next year should be your old conditional credence that John will die in the next year, conditional on his being a healthy skydiver (rather than conditional only on his being healthy, or only on his being a skydiver). But the Requirement of Total Evidence is in principle separable from the Rule of Conditionalization; even if some other update rule is correct, we still might want to insist that rational agents use a total statement of their new evidence as an input to that update rule.

Though the application of the RTE to the skydiver case is clear enough, there are some cases where its applicability is less clear. Suppose that I'm trying to assess your ability to throw a dart into the bullseye of a standard dartboard, and suppose that you hit point p inside the bullseye. It's at least somewhat natural to think that, when I update my beliefs about your dart-throwing ability, what is relevant is just the fact that your dart hit *some point or other inside the bullseye*, not that it was *point p in particular* that your dart hit. Similarly, some approaches to statistics entail that, when some experimental outcome E is observed, what is relevant is how likely it was that *E or some outcome at least as extreme as E would occur*, according to the "null hypothesis" that is being tested.[38] These approaches seem to be in some tension with the RTE, since when we know that E occurred, "E occurred" is a more complete statement of our evidence than "E or some outcome at least as extreme as E occurred."

29.6 Evidential Probability

There are two distinct philosophical theories that go by the name "evidential probability," one due primarily to Henry Kyburg (Kyburg 1961, Kyburg and Teng 2001) and one due to Timothy Williamson (Williamson 1998 and 2000).

Kyburg's theory of evidential probability was based on the idea that probabilities should be determined by relative frequencies. Evidential probability is a kind of conditional probability; the evidential probability of a sentence χ is evaluated relative to a set of sentences Γ_δ, which represents background knowledge, including knowledge of the proportions of objects satisfying various "reference class" predicates which also satisfy various "target class" predicates. The evidential probability of χ given Γ_δ, $\text{Prob}(\chi, \Gamma_\delta)$, is interval-valued, in order to accommodate cases where Γ_δ does not specify precise statistical information about the proportion of objects satisfying some reference class predicate which also satisfy the target class predicate. For example, suppose that that the proportion

[38] See Howson and Urbach 1993 for an overview of "frequentist" approaches to statistical inference which have this feature.

of red balls in an urn is known only to be between 20% and 30% inclusive, and that ball b is drawn from the urn. Then, Γ_δ will include sentences expressing each of these facts, and if we do not know anything *further* of relevance about b or the urn, then Prob(Red(b), Γ_δ)=[.2,.3]. (In a case where Γ_δ specifies that *precisely* 25% of the balls in the urn are red, Prob(Red(b), Γ_δ)=[.25,.25].) Of course, an individual object might belong to several different reference classes, each of which might contain a different proportion of individuals satisfying a particular target class predicate; for example, a individual might be both a ball in the urn and also a *plastic* ball in the urn, and the known proportion of red balls among the balls in the urn might be different from the known proportion of red balls among the plastic balls in the urn. This generates an instance of the "problem of the reference class," which Kyburg developed a procedure for solving. In the example above, Kyburg's theory entails that Prob(Red(b), Γ_δ) is equal to the known proportion of red balls among the *plastic* balls in the urn (rather than the known proportion of red balls among the balls in the urn, if this proportion is different), since the known proportion of red balls among the plastic balls is a more *specific* statistical statement. In addition to this "Specificity" principle, Kyburg also defended principles of "Richness" and "Strength," which can be used to solve reference class problems in cases where Specificity alone doesn't provide a unique solution.[39]

In Williamson 1998 and 2000, Timothy Williamson developed a distinct notion of evidential probability. To characterize evidential probabilities, Williamson assumes an "initial probability distribution" P, which is a probability function in the sense articulated in Section 29.3.1 above, and which measures "something like the intrinsic plausibility of hypotheses prior to investigation; this notion of intrinsic plausibility can vary in extension between contexts."[40] On Williamson's view, the evidential probability of H on total evidence E is $P(H|E)$; in other words, it is the conditional probability that the P function assigns to H, conditional on E. Evidential probabilities are distinct from credences. On Williamson's understanding, the existence of an evidential probability for H, given E, of x does not entail that anyone's credence in H actually is x. After all, E might not in fact be anyone's total evidence; furthermore, present evidence might not count at all in favor of H, and yet everyone might still be irrationally certain of H. Similarly, evidential probabilities are distinct from objective physical chances or objective frequencies; a law of nature, for instance, might have an objective physical chance of 1 even though present evidence tells against its truth.

One consequence of this view is that any agent's total evidence itself has evidential probability 1 for her; $P(E|E) = 1$ whenever it is defined. In this sense, an agent's total evidence is certain for her. However, while the assumption that the agent updates only by the Rule of Conditionalization entails that any proposition to which an agent attaches credence 1 retains a credence of 1 forever,[41] Williamson is interested to avoid the result that a proposition that has evidential probability 1 for an agent at a time (such as the agent's total evidence at that time) must *retain* evidential probability 1 for her permanently. One reason

[39] See Wheeler and J. Williamson 2011 for a complete discussion of Kyburg's theory.

[40] Williamson 2000, p. 210.

[41] Suppose that $p(H) = 1$. Since $p(H|E) = \frac{p(H \wedge E)}{p(E)}$ and since $p(H) = 1$ entails that $p(H \wedge E) = p(E)$, it follows that $p(H|E) = \frac{p(E)}{p(E)} = 1$; thus, an agent who updates only by the Rule of Conditionalization will retain a credence of 1 in H, regardless of what evidence she acquires.

for this is that an agent might *forget* some proposition; though it was certain for her at some earlier time t_1, it is no longer certain for her at a later time t_2.

Let P be the initial probability distribution, let E_α be the the total evidence of someone in situation α, and let $P_\alpha(H)$ be the evidential probability of H for someone in α. On Williamson's view, $P_\alpha(H) = P(H|E_\alpha) = P(H \wedge E_\alpha)/P(E_\alpha)$, where $P(E_\alpha) > 0$. This allows the evidential probability of E_α to decrease from 1; in situation α, the agent's evidential probability for E_α is 1, but if the agent forgets something and her situation changes from α to $\alpha*$, then $E_\alpha*$ need not entail E_α, so it is possible that $P_{\alpha*}(E_\alpha) < 1$ even though $P_\alpha(E_\alpha) = 1$.

Here are two questions that might reasonably be asked about Williamsonian evidential probabilities. First: Do they assume the Uniqueness Thesis, according to which there is a unique rational credence to assign to a given proposition on the basis of a fixed body of total evidence? Though Williamson's use of the construction "*the* evidential probability of H on E" suggests that it might, it seems as though this assumption could be relaxed; if there is range of rationally permissible responses to a particular body of evidence, it's not obvious that there is anything in Williamson's theory that prevents the evidential probability of H on E from taking on a range of values. Second: Is something like the Principal Principle (see Section 29.3.4) supposed to hold for evidential probabilities? In other words, suppose that X is the proposition that the *evidential probability* (rather than the *objective chance*) of A on evidence E is x; is it plausible that the agent's conditional probability for A on X and E, $p(A|X \wedge E)$, should be equal to x? Williamson doesn't address this question explicitly, but it seems quite plausible that he does understand evidential probabilities to obey this principle; after all, evidential probability is supposed to correspond to an objective notion of the probability conferred on some hypothesis by a given body of evidence, and so it seems as though a rational agent should assign a conditional credence to A, conditional on E and the information that the E confers an evidential probability of x on A, of x. Note that, unlike the case of the original Principal Principle, there can't be any such thing as inadmissible evidence here; any evidence that an agent had would be part of E.

Williamson's theory of evidential probability is very different from Kyburg's; whereas Kyburg was interested to derive evidential probabilities from statistical information alone, Williamson's theory applies to *any* body of total evidence. Both theories are also distinct from Carnap's theory of logical probabilities. Carnap's theory deserves the name "logical probability" because he thinks of probabilities as deriving from *syntactic* features of various descriptions of the domain; this is not a feature of Williamson's theory. In Carnap 1950 he assumes a domain of individuals and a number of monadic predicates that these individuals may or may not satisfy. A state description is a maximally specific description which settles whether or not each of the individuals in the domain satisfies each of the predicates under consideration. State descriptions are then grouped into equivalence classes (called "structure descriptions") under relabelings of the individuals in the domain; for example, since the state description which entails that a satisfies predicate F but that neither b nor c do can be "relabeled" to form the state description according to which b satisfies F but neither a nor c do, these two state descriptions belong to the same structure description (i.e., the structure description which says that exactly one thing is F). According to Carnap's theory, equal probability should be assigned to each structure description, and the various state descriptions that correspond to a particular structure description should be assigned an equal share of the probability assigned to that structure description. Thus, homogeneous state descriptions (such as "All of a, b, and c have F" or "None of a, b, or c has F") get

higher probabilities than non-homogeneous state descriptions (such as "*a* and *b* have *F*, but *c* doesn't have *F*"), since homogeneous structure descriptions are compatible with fewer state descriptions than non-homogeneous structure descriptions are. One consequence of this, for Carnap, is that the observation of an individual satisfying *F* constitutes evidence for the homogeneous state description according to which everything is *F*. (Carnap later dropped the assumption that there is one uniquely correct inductive method, and instead proposed his "continuum of inductive methods," each member of which corresponds to learning from experience at a different "rate." See Carnap 1952 and 1963b, and Sandy Zabell's chapter "Symmetry Arguments in Probability" (15) in this volume (Zabell 2016).)

What unites these three distinct theories is the goal of characterizing an objective notion of probability that corresponds to how likely some hypothesis is *on a given body of evidence*, even though each theory understands the nature of that "body of evidence" differently.

29.7 Sharp and Fuzzy Credences

As discussed in Section 29.3.1, Probabilism entails that a rational agent's credal state is modeled by a real-valued probability function. One worry that we might have about this feature of Probabilism is that it seems psychologically unrealistic; it is hard to see what could make it the case that my credence that (say) my dog will live past 10 years old is .549123425, rather than .549123426 or .549123424. Also, it has struck some philosophers as implausible that every set of total evidence *E* rationally mandates one particular credence function; sometimes, evidence seems to be equivocal in a way that fails to rationally constrain an agent to have one particular credal state. One possible response would be to adopt a "permissive" epistemology, according to which an agent with total evidence *E* is (at least sometimes) rationally permitted to have any of multiple different credence functions.[42] A different response is to relax the requirement that a rational agent's credal state be modeled by a *single* probability function, and to allow instead that an agent's credal state be modeled by a *set* of probability functions, called a Representor.[43] If there is more than one probability function in an agent's Representor, and those functions assign different values to a proposition *A*, then an agent's credence in *A* can't be expressed as a single value; rather, the agent's credence in *A* is represented as the set of all of the values that at least one of the functions in his Representor assigns to *A*. Thus, on this view, my total evidence might permit (or even require) me to have a credence in the proposition that my dog will live past 10 years old that is specified by the range [.5,.6]. In such a situation, say that my credence in *A* is "fuzzy."[44]

Despite the appeal of fuzzy credences, there are some worries that we might have about them. First, if it is psychologically unrealistic to think that I have a precise point-valued credence in the proposition that my dog will live past 10 years old, it seems equally

[42] Roger White explains this view in White 2005, though he ultimately rejects it.

[43] For a more detailed summary of this sort of strategy, see Fabio Cozman's chapter "Imprecise Probabilities" (14) in this volume (Cozman 2016).

[44] Various terms have been associated with this phenomenon. White 2010 refers to such credences as "mushy," Joyce 2005 as "indefinite," van Fraassen 1990 as "vague," Levi 1974 as "indeterminate," Walley 1991 as "imprecise," and Sturgeon 2008 as "thick."

psychologically unrealistic to think that I have a fuzzy credence in that proposition that is represented by precisely the interval [.5,.6], rather than the interval [.500001,.600001] or [.499999,.599999]; thus, some sort of theory of "higher-order" fuzziness seems to be required. Secondly, White 2010 argues that there is a tension between the rational permissibility of fuzzy credences and the Reflection principle (see Section 29.3.3); he considers a case where (on the assumption that fuzzy credences are rationally permissible) it seems as though I'm certain that my credence in *A* will rationally be fuzzy in the future, and yet where it's implausible that my credence in *A* rationally should be fuzzy now.[45] Thirdly, Elga 2010 argues that, regardless of his credence in *H*, any rational agent should be disposed to accept at least one of the following bets:

Bet A: If *H* is true, you lose $10. Otherwise you win $15.
Bet B: If *H* is true, you win $15. Otherwise you lose $10.

But, Elga argues, an agent with a fuzzy credence in *H* might be rationally permitted to reject both bets; if the range of his credence in *H* is sufficiently spread out, his credence in *H* may permit him to reject both Bet A and Bet B, and thereby forgo the guaranteed $5 that would result from accepting both bets.

29.8 Likelihood Arguments

Likelihood Arguments are arguments of a particular style that have been applied in various places in epistemology. Likelihood Arguments proceed by appealing to the Likelihood Principle, which says *E* supports $H1$ over $H2$ just in case $H1$ makes *E* likelier than $H2$ does. Suppose, for instance, that $H1$ is the proposition that the lottery is fair, $H2$ is the proposition that the lottery is heavily rigged in Mary's favor, and *E* is the proposition that Mary wins the lottery three times in a row. $H1$ seems to make *E* quite unlikely; a fair lottery is very unlikely to result in three wins in a row for anyone. But $H2$ looks to make *E* rather more likely; if the lottery is heavily rigged in Mary's favor, then it wouldn't be at all surprising for Mary to win three times. According to the Likelihood Principle, it follows that Mary's three wins are evidence for the hypothesis of rigging over the hypothesis of fairness.

The Fine-Tuning Argument is one instance of the Likelihood Argument strategy. According to the Fine-Tuning Argument, if the fundamental constants of the universe were set by chance or some other mindless natural process, it would have been very unlikely that they would all just happen to take on values that are in the narrow "life-sustaining range" and hence that can support life.[46] By contrast, if the fundamental constants were set by an Intelligent Designer who was interested in seeing to it that the universe is capable of supporting life, then it would have been far more likely that the constants would take on

[45] White 2010, pp. 175–7.

[46] For a general philosophical overview of Fine-Tuning Arguments, see Leslie 1989, Holder 2004, Sober 2004, and the essays in Manson 2003. For a discussion of the physical details of Fine-Tuning Arguments, see Barrow and Tipler 1986, Collins 2003, Holder 2004 chapter 3, and Ellis 2006 section 9.1. For some objections to Fine-Tuning Arguments, see McGrew et al. 2005, Narveson 2003, and Sober 2009.

values in the life-sustaining range. Thus, according to the Likelihood Principle, the fact that the actual constants have values in the life-sustaining range is evidence for the hypothesis of Design over the hypothesis of Chance. Likelihood Arguments have also been applied to the related question of whether the fact that the constants are life-sustaining is evidence for the Multiple Universe Hypothesis, according to which there are many non-interacting universes, each with its own set of fundamental constants. White 2003 argues that it is not; though the Multiple Universe Hypothesis does make it likelier than the Single Universe Hypothesis makes it that there would be *some* life-friendly universe, it does not make it likelier that *our* universe would be life-friendly. Bradley 2009 and Manson and Thrush 2003 develop the opposite view.

29.9 Dogmatism

In Pryor 2000, James Pryor developed an influential account of perceptual justification that he calls "dogmatism." According to dogmatism, an agent's perceptual experiences can give him prima facie justification to believe the basic contents of those experiences, even if he doesn't have any "antecedent" justification to believe that his perceptual faculty is reliable, or that he isn't a brain-in-a-vat, etc.; all that is required, according to dogmatism, is that the agent *lack* any antecedent justification to believe that such hypotheses are false. So, according to dogmatism, even if I lack justification to believe (say) that I'm not a handless brain-in-a-vat, an experience as of a hand in front of me can give me prima facie justification to believe that I have hands; and if that prima facie justification is undefeated, then my experience as of a hand can give me all-things-considered justification to believe that I have hands.

Roger White has raised the following objection to dogmatism.[47] Call the hypothesis that I am a handless brain-in-a-vat being stimulated to have hand-experiences S, and call the hypothesis that I have hands H. Suppose that, at t_1, I'm not justified in believing $\neg S$. According to dogmatism, as long as I'm also not justified at t_1 in believing S, my hand-experience E can justify me in believing H. Suppose that it does, and consider some time, t_2, after I've had E, at which I'm justified in believing H. Since I know that H entails $\neg S$, and I'm justified at t_2 in believing H, I should be able to competently deduce $\neg S$ from H, thereby becoming justified in believing $\neg S$ at t_2 too. But, White argues, this is quite odd. The only thing that has happened between t_1 and t_2 is that I have had experience E. But the proposition that I have E is *entailed* by S; if I am a handless brain-in-a-vat being stimulated to have hand-experiences, then I am certain to have E. And, White continues, it's very implausible that an experience entailed by S can take me from a situation at t_1 where I am not justified in believing $\neg S$, to a situation at t_2 where I am justified in believing $\neg S$. After all, S "said" that E would occur, so E's occurrence can't count *against* S by justifying a belief in $\neg S$. One way of supporting this point is by appealing to the Likelihood Principle discussed in Section 29.7. If S entails E, then S makes E maximally likely; but if (as the Likelihood Principle entails) E is evidence for $\neg S$ over S just in case $\neg S$ makes E likelier

[47] See White 2006. Stephen Schiffer makes a related, but distinct, objection in Schiffer 2004.

than S does, then E can't be evidence for $\neg S$ over S, since no hypothesis could make E any likelier than S does (and a fortiori $\neg S$ can't). But if E isn't evidence for $\neg S$ over S, then it is hard to see how I could be justified in believing $\neg S$ at t_2, given that I wasn't justified in believing $\neg S$ at t_1. But, according to the dogmatist, this scenario is possible.

Various responses to White's objection have been pursued. It has been generally agreed that White's objection is successful in demonstrating that dogmatism is incompatible with a probabilistic model of credences, so the responses on behalf of dogmatism have placed the blame for this incompatibility on that probabilistic model, rather than dogmatism. Pryor himself has proposed a non-probabilistic framework for credences which is based on Dempster-Shafer functions and which is designed to avoid White's objection.[48] Weatherson 2007 argues that White's objection can be avoided by a "fuzzy" model of credences of the sort discussed in Section 29.6.

29.10 Transmission

In various places, Crispin Wright, Martin Davies, James Pryor, and others have discussed the question of whether various arguments exhibit the phenomenon of *transmission-failure*.[49] Consider some evidence E, a hypothesis $H1$, and another hypothesis $H2$, and suppose that the relevant agent knows that $H1$ entails $H2$. Wright argues that in some cases, E provides an epistemic warrant for $H1$ and, in virtue of the fact that the agent knows that $H1$ entails $H2$, E thereby provides an epistemic warrant for $H2$. In such cases, we say that there is transmission-success. Wright (2000, 2003) thinks that the following case exhibits transmission-success:

Zebra
E: My experience is as of a zebra in a pen in front of me.
$H1$: There is a zebra in a pen in front of me.
$H2$: There is an animal in a pen in front of me

In Zebra, Wright 1985 claims, E can provide the agent with a warrant to believe $H1$, which can then "transmit" through the known entailment from $H1$ to $H2$, and thereby also provide the agent with a warrant to believe $H2$.

By contrast, Wright claims, some arguments exhibit transmission-failure. Consider:

Zebra*
E: My experience is as of a zebra in a pen in front of me.
$H1$: There is a zebra in a pen in front of me.
$H2^*$: It is not the case that there is a mule cleverly disguised to look like a zebra in a pen in front of me.

[48] Pryor sets out this framework in Pryor 2007a, and provides some further philosophical motivation for the framework in Roberts 2001. For an excellent overview of Dempster-Shafer Theory, see Weisberg 2011 section 1.2.3 and Weisberg 2010.

[49] Wright 2000, Wright 2002, Wright 2003, Wright 2004, Davies 2003, Davies 2004, Pryor 2004, Pryor 2012b.

Wright argues that in ZEBRA*, even though E does provide a warrant for the agent to believe $H1$, and even though the agent knows that $H1$ entails $H2$*, E does *not* provide a warrant for the agent to believe $H2$*. The reason, Wright argues, is that in ZEBRA*, E provides a warrant to believe $H1$ only because the subject *already* had *independent* warrant to believe $H2$*; thus, E cannot provide any *new* warrant to believe $H2$*. By contrast, in ZEBRA, since the warrant that E provides the agent to believe $H1$ does *not* depend on her already having independent warrant to believe $H2$, the warrant that E provides for $H1$ *is* able to "transmit" through the known entailment, and thereby provide the agent with warrant to believe $H2$.

There have been a variety of attempts to translate Wright's discussion—as well as his strategy for distinguishing arguments exhibiting transmission-success from those exhibiting transmission-failure—into probabilistic terms. One such attempt is due to Okasha 2004. In cases of transmission-failure, Okasha argues, background conditions make it such that E does provide evidence for $H1$ on the assumption that $H2$ is true, so $p(H1|E \wedge H2) > p(H1|H2)$. But, since the evidence that E provides for $H1$ depends on the agent's having independent reason to believe $H2$, background conditions make it such that E does *not* provide evidence for $H1$ if we do not assume that $H2$ is true, so $p(H1|E) \leq p(H1)$. Okasha then shows that, together with the assumption that $H1$ entails $H2$, these conditions entail that $p(H2|E) \leq p(H2)$; in other words, E does not provide any evidence for $H2$.

Chandler 2010 and Moretti 2012 raise various objections to Okasha's proposal, and propose probabilistic reconstructions of their own of Wright's argument. Other proposals are developed in Kotzen 2012 and Moretti and Piazza 2013. Tucker 2010 develops a view which entails that ZEBRA* does exhibit transmission-success after all.

References

Appiah, A. (1985) *Assertion and Conditionals*. Cambridge: Cambridge University Press.

Armendt, B. (1993) Dutch Book, Additivity, and Utility Theory. *Philosophical Topics*. 21. pp. 1–20.

Arntzenius, F. (2002) Reflections on Sleeping Beauty. *Analysis*. 62. pp. 53–62.

Arntzenius, F. (2003) Some Problems for Conditionalization and Reflection. *Journal of Philosophy*. 100. 7. pp. 356–70.

Barrow, J. and Tipler, F. (1986) *The Anthropic Cosmological Principle*. New York, NY: Oxford University Press.

Bernstein, A. and Wattenberg, F. (1969) Non-standard Measure Theory. In Luxemburg, W. (ed.) *Applications of Model Theory of Algebra, Analysis, and Probability*. New York, NY: Holt, Rinehard and Winston.

Bradley, D. (2009) Multiple Universes and Observation Selection Effects. *American Philosophical Quarterly*. 46. pp. 61–72.

Bronfman, A. (ms.) A Gap in Joyce's Argument for Probabilism. Unpublished manuscript.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Carnap, R. (1952) *The Continuum of Inductive Methods*. Chicago, IL: University of Chicago Press.

Carnap, R. (1963a) Carnap's Intellectual Autobiography. In Schilpp, P. A. (ed.) *The Philosophy of Rudolf Carnap. The Library of Living Philosophers*. Vol. XI. La Salle, IL: Open Court.

Carnap, R. (1963b) Replies and Systematic Expositions. In Schilpp, P. A. (ed.) *The Philosophy of Rudolf Carnap. The Library of Living Philosophers*. Vol. XI. La Salle, IL: Open Court.

Chandler, J. (2010) The Transmission of Support: A Bayesian Re-analysis. *Synthese*. 176. pp. 333–43.
Christensen, D. (1991) Clever Bookies and Coherent Beliefs. *The Philosophical Review*. 100. pp. 229–47.
Christensen, D. (1996) Dutch-Book Arguments Depragmatized: Epistemic Consistency for Partial Believers. *Journal of Philosophy*. 93. pp. 450–79.
Christensen, D. (2004) *Putting Logic in Its Place: Formal Constraints on Rational Belief*. Oxford: Oxford University Press.
Collins, R. (2003) Evidence for Fine-Tuning. In Manson (2003). pp. 178–99.
Cozman, F. (2016) Imprecise and Indeterminate Probabilities. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy.* Oxford: Oxford University Press.
Davies, M. (2003) The Problem of Armchair Knowledge. In Nuccetelli, S. (ed.) *New Essays on Semantic Externalism and Self-Knowledge*. pp. 23–156. Cambridge, MA: MIT Press.
Davies, M. (2004) Epistemic Entitlement, Warrant Transmission and Easy Knowledge. *Proceedings of the Aristotelian Society*. Supplementary Volume 78. pp. 213–45.
de Finetti, B. (1937) La Prévision: Ses Lois Logiques, Ses Sources Subjectives. *Annales de l'Institut Henri Poincaré*. 7. pp. 1–68.
Dorr, C. (2002) Sleeping Beauty: In Defence of Elga. *Analysis*. 62. pp. 292–6.
Easwaran, K. (2011) Bayesianism I: Introduction and Arguments in Favor. *Philosophy Compass*. 6. 5. pp. 312–20.
Edwards, W., Lindman, H., and Savage, L. J. (1963) Bayesian Statistical Inference for Psychological Research. *Psychological Review*. 70. pp. 193–242.
Elga, A. (2000) Self-locating Belief and the Sleeping Beauty Problem. *Analysis*. 60. 143–7.
Elga, A. (2010) Subjective Probabilities Should be Sharp. *Philosophers' Imprint*. 10. 5.
Ellis, G. (2006) Issues in the Philosophy of Cosmology. [Online] Available from: http://arxiv.org/abs/astro-ph/0602280v2. [Accessed 1 Oct 2015.]
Foley, R. (1993) *Working Without a Net*. Oxford: Oxford University Press.
Frankish, K. (2009) Partial Belief and Flat-out Belief. In Huber, F. and Schmidt-Petri, C. (eds.) *Degrees of Belief*. pp. 75–93. Berlin: Springer.
Greaves, H. and Wallace, D. (2006) Justifying Conditionalization: Conditionalization Maximizes Expected Epistemic Utility. *Mind*. 115. pp. 607–32.
Hájek, A. (2003) What Conditional Probabilities Could Not Be. *Synthese*. 137. pp. 273–323.
Hájek, A. (2008) Arguments for—or Against—Probabilism. *The British Journal for the Philosophy of Science*. 59. 4. pp. 793–819.
Hájek, A. (2012) Is Strict Coherence Coherent? *Dialectica*. 66. 3. pp. 411–24.
Hájek, A. (unpublished A) A Puzzle about Partial Belief.
Hájek, A. (unpublished B) Staying Regular.
Hall, N. (1994). Correcting the Guide to Objective Chance. *Mind*. 103. pp. 505–17.
Halpin, J. (1994) Legitimizing Chance: The Best-System Approach to Probabilistic Laws in Physical Theory. *Australasian Journal of Philosophy*. 72. pp. 317–38.
Hitchcock, C. (2004) Beauty and the Bets. *Synthese*. 139. pp. 405–20.
Holder, R. (2004) *God, the Multiverse, and Everything: Modern Cosmology and the Argument from Design*. Burlington, VT: Ashgate.
Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach*. 2nd Edition. Chicago: Open Court.
Hunter, D. (1996) On the Relation Between Categorical and Probabilistic Belief. *Noûs*. 30. pp. 75–98.

Jackson, F. (1987) *Conditionals*. Oxford: Blackwell.
Jeffrey, R. (1965) *The Logic of Decision*. Chicago, IL: University of Chicago Press.
Jeffrey, R. (1970) Dracula meets Wolfman: Acceptance vs. Partial Belief. In Swain, M. (ed.) *Induction, Acceptance, and Rational Belief*. pp. 157–85. Dordrecht: D. Reidel Publishing Company.
Jeffrey, R. (1992) *Probability and the Art of Judgment*. Cambridge: Cambridge University Press.
Jeffreys, H. (1961) *Theory of Probability*. 3rd Edition. Oxford: Clarendon Press.
Joyce, J. (1998) A Non-Pragmatic Vindication of Probabilism. *Philosophy of Science*. 65. pp. 575–603.
Joyce, J. (1999) *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
Joyce, J. (2004) Bayesianism. In Mele, A. R. and Rawlings, P. (eds.) *The Oxford Handbook of Rationality*. pp. 132–55. Oxford: Oxford University Press.
Joyce, J. (2005) How Probabilities Reflect Evidence. *Philosophical Perspectives*. 19. pp. 153–78.
Joyce, J. (2009) Accuracy and Coherence: Prospects for an Alethic Epistemology of Partial Belief. In Huber, F. and Schmidt-Petri, C. (eds.) *Degrees of Belief*. Dordrecht: Springer.
Kaplan, M. (1996) *Decision Theory as Philosophy*. Cambridge: Cambridge University Press.
Kemeny, J. (1955) Fair Bets and Inductive Probabilities. *Journal of Symbolic Logic*. 20. pp. 263–73.
Kolmogorov, A. (1956) *Foundations of the Theory of Probability*. 2nd Edition. New York: Chelsea Publishing Company.
Kotzen, M. (2012) Dragging and Confirming. *The Philosophical Review*. 121. pp. 55–93.
Kyburg, H. (1961) *Probability and the Logic of Rational Belief*. Middleton, CT: Wesleyan University Press.
Kyburg, H. and Teng, C. (2001) *Uncertain Inference*. Cambridge: Cambridge University Press.
Leslie, J. (1989) *Universes*. New York, NY: Routledge.
Levi, I. (1974) On Indeterminate Probabilities. *Journal of Philosophy*. 71. pp. 391–418.
Lewis, D. (1979) Attitudes De Dicto and De Se. *The Philosophical Review*. 88. pp. 513–43.
Lewis, D. (1980) A Subjectivist's Guide to Objective Chance. In Jeffrey, R. C. (ed.) *Studies in Inductive Logic and Probability*. Vol. 2, pp. 263–93. Berkeley, CA: University of California Press.
Lewis, D. (1986) Introduction. In Lewis, D. *Philosophical Papers*, Vol. 2, pp. ix–xvii. Oxford: Oxford University Press.
Lewis, D. (1994) Chance and Credence: Humean Supervenience Debugged. *Mind*. 103. pp. 473–90.
Lewis, D. (1999) Why Conditionalize? In Lewis, D. *Papers in Metaphysics and Epistemology*. pp. 403–7. Cambridge: Cambridge University Press.
Lewis, D. (2001) Sleeping Beauty: Reply to Elga. *Analysis*. 61. pp. 171–6.
Loewer, B. (2004) David Lewis's Humean Theory of Objective Chance. *Philosophy of Science*. 71. pp. 1115–25.
Maher, P. (1993) *Betting on Theories*. Cambridge: Cambridge University Press.
Maher, P. (2002) Joyce's Argument for Probabilism. *Philosophy of Science*. 69. pp. 73–81.
Manson, N. (ed.) (2003) *God and Design: The Teleological Argument and Modern Science*. New York, NY: Routledge.
Manson, N. and Thrush, M. (2003) Fine-Tuning, Multiple Universes, and the 'This Universe' Objection. *Pacific Philosophical Quarterly*. 84. pp. 67–83.

McGrew, T., McGrew, L., and Vestrup, E. (2001) Probabilities and the Fine-Tuning Argument: A Sceptical View. *Mind*. 110. 440. pp. 1027–37

Meacham, C. (2008) Sleeping Beauty and the Dynamics of De Se Beliefs. *Philosophical Studies.* 138. pp. 245–69.

Mellor, D. H. (1971) *The Matter of Chance*. Cambridge: Cambridge University Press.

Moretti, L. (2012) Wright, Okasha and Chandler on Transmission Failure. *Synthese*. 184. pp. 217–34.

Moretti, L. and Piazza, T. (2013) When Warrant Transmits and When It Doesn't: Towards a General Framework. *Synthese*. 190. 13. pp. 2481–503.

Moss, S. (2013) Epistemology Formalized. *The Philosophical Review.* 122. 1. pp. 1–43.

Narveson, J. (2003) God by Design? In Manson, N. *God and Design: The Teleological Argument and Modern Science*. New York, NY: Routledge. pp. 88–104.

Okasha, S. (2004) Wright on the Transmission of Support: A Bayesian Analysis. *Analysis*. 64. pp. 139–46.

Perry, J. (1979) The Problem of the Essential Indexical. *Noûs.* 13. pp. 3–21.

Pryor, J. (2000) The Skeptic and the Dogmatist. *Noûs.* 34. pp. 517–49.

Pryor, J. (2004) What's Wrong with Moore's Argument? *Philosophical Issues.* 14. pp. 349–78.

Pryor, J. (2007a) What's Wrong with McKinsey-Style Reasoning? In Goldberg, S. (ed.) *Internalism and Externalism in Semantics and Epistemology* Oxford: Oxford University Press.

Pryor, J. (2007b) Uncertainty and Undermining. [Online] Available from: http://www.jimpryor.net/research/papers/Uncertainty.pdf. [Accessed 1 Oct 2015]

Pryor, J. (2012a) Problems for credulism [Online] Available from: http://www.jimpryor.net/research/papers/Credulism.pdf. [Accessed 1 Oct 2015]

Pryor, J. (2012b) When Warrant Transmits. In Coliva, A. (ed.) *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright*. pp. 269–303. Oxford: Oxford University Press.

Quine, W. V. (1969) Propositional Objects. In *Ontological Relativity and Other Essays*. New York, NY: Columbia University Press.

Ramsey, F. (1926) Truth and Probability. In Braithwaite, R. B. (ed.) *The Foundations of Mathematics and Other Logical Essays*, pp. 156–98. London: Kegan, Paul, Trench, Trubner & Co.; New York, NY: Harcourt, Brace and Company.

Roberts, J. (2001) Undermining Undermined: Why Humean Supervenience Never Needed to Be Debugged (Even If It's a Necessary Truth). *Philosophy of Science*, 68. 3. Supplement Proceedings of the 2000 Biennial Meeting of the Philosophy of Science Association. Part I: Contributed Papers (Sep., 2001) pp. S98–S108.

Savage, L. (1954) *The Foundations of Statistics*. New York, NY: John Wiley and Sons.

Schick, F. (2003) *Ambiguity and Logic*. Cambridge: Cambridge University Press.

Schiffer, S. (2004) Skepticism and the Vagaries of Justified Belief. *Philosophical Studies.* 119. pp. 161–84.

Schwarz, W. (2016) Best Systems Approaches to Chance. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

Seidenfeld, T. (1985) Calibration, Coherence, and Scoring Rules. *Philosophy of Science.* 52. pp. 274–94.

Shimony, A. (1955) Coherence and the Axioms of Confirmation. *Journal of Symbolic Logic.* 20. pp. 1–28.

Shimony, A. (1970) Scientific Inference. In Colodny, R. (ed.) *The Nature and Function of Scientific Theories*. Pittsburgh, PA: University of Pittsburgh Press.
Shimony, A. (1988) An Adamite Derivation of the Calculus of Probability. In Fetzer, J. H. (ed.) *Probability and Causality*. pp. 151–61. Dordrecht: Reidel.
Skyrms, B. (1980) *Causal Necessity*. New Haven, CT: Yale University Press.
Skyrms, B. (1984) *Pragmatism and Empiricism*. New Haven, CT: Yale University Press.
Skyrms, B. (1995) Strict Coherence, Sigma Coherence, and the Metaphysics of Quantity. *Philosophical Studies*. 77.1. pp. 39–55.
Sober, E. (2004) The Design Argument. In Mann, W. E. (ed.) *Blackwell Guide to Philosophy of Religion*. pp. 117–47. New York, NY: Blackwell Publishers.
Sober, E. (2009) Absence Of Evidence And Evidence Of Absence: Evidential Transitivity In Connection With Fossils, Fishing, Fine-Tuning, and Firing Squads. *Philosophical Studies.* 143. pp. 63–90.
Stalnaker, R. (1970). Probability and Conditionals. *Philosophy of Science*. 37. pp. 64–80.
Stalnaker, R. (1981) Indexical Belief. *Synthese.* 49. pp. 129–51.
Stalnaker, R. (1984). *Inquiry*. Cambridge, MA: MIT Press.
Sturgeon, S. (2008) Reason and the Grain of Belief. *Noûs.* 42. pp. 139–65.
Talbott, W. (1991) Two Principles of Bayesian Epistemology. *Philosophical Studies.* 62. pp. 135–50.
Teller, P. (1973) Conditionalization and Observation. *Synthese.* 26. pp. 218–58.
Thau, M. (1994) Undermining and Admissibility. *Mind.* 103. pp. 491–503.
Titelbaum, M. (2016) Self-Locating Credences. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Tucker, C. (2010) When Transmission Fails. *Philosophical Review.* 119. pp. 497–529.
van Fraassen, B. (1983) Calibration: A Frequency Justification for Personal Probability. In Cohen, R. and Laudan, L. (eds.) *Physics Philosophy and Psychoanalysis*. pp. 295–319. Dordrecht: D. Reidel.
van Fraassen, B. (1984) Belief and the Will. *Journal of Philosophy.* 81. pp. 235–56.
van Fraassen, B. (1990) Figures in a Probability Landscape. In Dunn, J. and Gupta, A. (eds.) *Truth or Consequences*. Dordrecht: Kluwer.
van Fraassen, B. (1995) Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies.* 77. pp. 7–37.
Vranas, P. (1998) Who's Afraid of Undermining? Paper read at the Biennial Meeting of the Philosophy of Science Association.
Walley, P. (1991) *Statistical Reasoning with Imprecise Probabilities*. London: Chapman and Hall.
Weatherson, B. (2005) Can We Do Without Pragmatic Encroachment? *Philosophical Perspectives.* 19. 1. pp. 417–43.
Weatherson, B. (2007) The Bayesian and the Dogmatist. *Proceedings of the Aristotelian Society.* 107. pp. 169–85.
Weisberg, J. (2011) Varieties of Bayesianism. In Gabbay, D., Hartmann, S., and Woods, J. (eds.) *Handbook of the History of Logic*. Vol 10. Oxford: North Hollsand.
Weisberg, J. (2010) Dempster-Shafer Theory. [Online] Available from: http://www.utm.utoronto.ca/~weisber3/unpublished/NIP\%20-\%20DST.pdf. [Accessed 10 Oct 2015.]
Wheeler, G. and Williamson, J. (2011) Evidential Probability and Objective Bayesian Epistemology. In Bandyopadhyay, P. S. and Forster, M. R. (eds.) *Philosophy of Statistics*. (Handbook of the Philosophy of Science, Vol. 7). Oxford: Elsevier.

White, R. (2003) Fine-Tuning and Multiple Universes. In Manson, N. *God and Design: The Teleological Argument and Modern Science*. New York, NY: Routledge. pp. 229–50.
White, R. (2005) Epistemic Permissiveness. *Philosophical Perspectives.* 19. pp. 445–59.
White, R. (2006) Problems for Dogmatism. *Philosophical Studies.* 131. pp. 525–57.
White, R. (2010) Evidential Symmetry and Mushy Credence. *Oxford Studies in Epistemology*. Vol. 3. Oxford: Oxford University Press.
Williamson, T. (1998) Conditionalizing on Knowledge. *British Journal for the Philosophy of Science.* 49. pp. 89–121.
Williamson, T. (2000) *Knowledge and its Limits*. Oxford: Oxford University Press.
Williamson, T. (2007) How Probable Is an Infinite Sequence of Heads? *Analysis.* 67. pp. 173–80.
Williamson, J. (2010) *In Defence of Objective Bayesianism*. Oxford: Oxford University Press.
Wright, C. (1985) Facts and Certainty. *Proceedings of the British Academy.* 71. pp. 429–72.
Wright, C. (2000) Cogency and Question-Begging: Some Reflections on McKinsey's Paradox and Putnam's Proof. *Philosophical Issues.* 10. pp. 140–63.
Wright, C. (2002) (Anti-)Sceptics Simple and Subtle: G. E. Moore and John McDowell. *Philosophy and Phenomenological Research.* 65. pp. 330–48.
Wright, C. (2003) Some Reflections on the Acquisition of Warrant by Inference. In Nuccetelli, S. (ed.) *New Essays on Semantic Externalism and Self-Knowledge*. pp. 57–77. Cambridge, MA: MIT Press.
Wright, C. (2004) Warrant for Nothing (and Foundations for Free?) *Proceedings of the Aristotelian Society*. Supplementary Vol. 78. pp. 167–212.
Zabell, S. (2016) Symmetry Arguments in Probability. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Zynda, L. (2000) Representation Theorems and Realism about Degrees of Belief. *Philosophy of Science*. 67. 1. pp. 45–69.
Zynda, L. (2016) Subjective Probability. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

CHAPTER 30

CONFIRMATION THEORY

VINCENZO CRUPI AND KATYA TENTORI

In philosophy of science, formal epistemology, and related areas, *confirmation* has become a key technical term. Broadly speaking, confirmation has to do with how evidence affects the credibility of hypotheses, an issue that is crucial to human reasoning in a variety of domains, from scientific inquiry to medical diagnosis, legal argumentation, and beyond. In what follows, we will address *probabilistic* theories of confirmation. The case for tackling confirmation in a probabilistic framework is easily put. The connection between evidence and hypothesis is typically fraught with uncertainty, and probability is widely recognized as the formal representation of uncertainty that is best understood and motivated.[1] We will thus frame our discussion by positing a set $\boldsymbol{P}$ of probability functions representing possible states of belief concerning a domain described in a (finite) propositional language L. We will also denote as L_c the set of contingent formulae in L (namely, those expressing neither logical truths nor logical falsehoods), and we will have hypothesis h and evidence e belonging to L_c. Finally, $\boldsymbol{P}$ will be assumed to include all *regular* probability functions that can be defined over L (i.e., such that, for any $\alpha \in L_c$ and any $P \in \boldsymbol{P}$, $0 < P(\alpha) < 1$).[2]

[1] Although well-established, probabilistic confirmation theory has not always been popular, nor has it remained unchallenged even in recent times. For prominent critical voices, see Kelly and Glymour (2004) and Norton (2010). As regards earlier influential and non-probabilistic accounts of confirmation, one should mention at least Popper's (1959) notion of "corroboration" through bold successful predictions and Hempel's (1943) analysis of confirmation by instances. There also exist cases which tend to defy the distinction between advocates and critics of probabilistic confirmation theory: Isaac Levi's work is a major example (e.g., Levi 2010). Finally, there are authors who rely on probability to account for evidential reasoning, but not as a representation of belief under uncertainty (as is the case throughout this chapter). This applies, for instance, to Royall's (1997) likelihoodism, as well as to Mayo's (1996) error-theoretic approach. Also see Crupi (2015) for a more extensive discussion.

[2] Regularity can be motivated as a way to represent credences that are non-dogmatic (see Howson 2000: p. 70). It is a very convenient assumption, but not an entirely innocent one. Festa (1999) and Kuipers (2000) discuss some limiting cases that are left aside here owing to this constraint.

30.1 Qualitative Confirmation Absolute vs. Incremental

A *qualitative* account of confirmation amounts to spelling out the conditions on which evidence *e* does or does not confirm hypothesis *h*. On the qualitative level of analysis, a clear distinction must be drawn between so-called *absolute* and *incremental* confirmation (see, e.g., Hájek and Joyce 2008). Adapting a useful piece of formalism (originally due to Gabbay 1982, and now standard in non-monotonic logics), we will employ "$\mathrel{|\!\sim}^{A}_{P}$" for "confirms in the absolute sense (relative to *P*)" and "$\mathrel{|\!\sim}^{I}_{P}$" for "confirms in the incremental sense (relative to *P*)".

(Abs) *Absolute confirmation*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $e \mathrel{|\!\sim}^{A}_{P} h$ if and only if $P(h|e) > r$ (with $\frac{1}{2} \leq r$).

(Inc) *Incremental confirmation*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $e \mathrel{|\!\sim}^{I}_{P} h$ if and only if $P(h|e) > P(h)$.

Absolute confirmation, as defined above, concerns whether or not the probability of *h* given *e* is *high enough* relative to a threshold value *r*. (This value must be separately specified and can set up a more or less demanding criterion.) On the other hand, incremental confirmation concerns whether or not the probability of *h* is *increased* when *e* is acquired as evidence. Before presenting and discussing critical divergences between them, let us point out that absolute and incremental confirmation share some rather basic properties. There follows a list of four, all of which are implied by each of both (Abs) and (Inc). Since they hold for absolute and incremental confirmation alike, superscripts *A* and *I* are removed.[3] (Also, notation is as expected, in that we write $\{\alpha, \beta\} \mathrel{|\!\sim}_P \gamma$ if and only if $(\alpha \wedge \beta) \mathrel{|\!\sim}_P \gamma$).

(EC) *Entailment condition*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, if $e \models h$, then $e \mathrel{|\!\sim}_P h$.

(NM) *Non-monotonicity*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$ such that $e \mathrel{|\!\sim}_P h$ and $e \not\models h$, there exists $x \in L_c$ such that $\{e,x\} \not\mathrel{|\!\sim}_P h$.[4]

(Cases) *Proof by cases*
For any $h,e,x \in L_c$ and any $P \in \boldsymbol{P}$, if $\{e,x\} \mathrel{|\!\sim}_P h$ and $\{e,\neg x\} \mathrel{|\!\sim}_P h$, then $e \mathrel{|\!\sim}_P h$.

(CComp) *Confirmation complementarity* (*qualitative*)

For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, if $e \mathrel{|\!\sim}_P h$ then $e \not\mathrel{|\!\sim}_P \neg h$.

[3] (EC) has been standard ever since Hempel (1945: p. 103) and it is analoguous to so-called *superclassicality* in logical parlance (see, e.g., Antonelli 2012). (NM) is inspired by Fitelson and Hawthorne (2010: p. 209). See Malinowski (2005) and Kuipers (2007) for earlier appearances of (Cases), and Crupi, Festa, and Buttasi (2010: p. 85) for remarks and terminology relevant to (CComp).

[4] The easiest way to prove (NM) is to just posit $x = \neg h$.

All four of the above conditions seem compelling upon reflection. First, relations of plain deductive entailment are instances of confirmation, as stated by (EC). Here, of course, hypothesis h is *conclusively* established in light of the evidence e, so that these are "ideal" and special instances, as it were. Indeed, confirmation is otherwise a form of non-monotonic reasoning in the sense of (NM), thus *non*-conclusive and defeasible. This is as it should be, motivated by the consideration that a hypothesis (say, Newtonian physics) can receive spectacular confirmation and nevertheless be overthrown in light of subsequent further evidence. However, as (Cases) implies, if confirmation of h happens not to be defeated by conjoining e to either of the statements x or $\neg x$, then e confirms h regardless. Finally, the claim that some evidence e confirms both hypothesis h and its negation $\neg h$ would be unintelligible, so that (CComp) also seems an obvious requirement.

Despite these preliminary remarks, it is important to realize that absolute and incremental confirmation convey very different concepts. Indeed, the distinction between the two – "extremely fundamental" and yet "sometimes unnoticed", as Salmon (1969: pp. 48–9) put it – has proved recurrently necessary for theoretical clarity (see, e.g., Crupi, Fitelson, and Tentori 2008). The following distinctive properties of $\vdash_P^A$ (both reaching back to Hempel 1945: pp. 103 ff.) will help us develop this point more thoroughly.

(SC) *Special consequence condition*
For any h_1,h_2,$e \in L_c$ and any $P \in \boldsymbol{P}$, if $h_1 \models h_2$ and $e \vdash_P^A h_1$, then $e \vdash_P^A h_2$.

(CC) *Consistency condition*
For any h_1,h_2,$e \in L_c$ and any $P \in \boldsymbol{P}$, if $\models \neg(h_1 \wedge h_2)$ and $e \vdash_P^A h_1$, then $e \nvdash_P^A h_2$.

In most contexts, "confirming evidence" is taken to be evidence which "makes a difference," to some extent at least, in favor of the hypothesis of interest. Bearing this in mind, it is then easy to show that (SC) is too inclusive, while (CC) is too restrictive. As a consequence, although a formally unobjectionable and historically influential notion, $\vdash_P^A$ does not seem to characterize confirmation very effectively. Let us discuss this line of argument in more detail.

As to the assessment of (SC), a simple numerical example will best serve our purposes. Suppose that a card is drawn from a well shuffled standard deck. Let h_1 be "the card drawn is a red non-face card" and let h_2 be "the card drawn is a non-face card," so that h_1 implies h_2. If the evidence e is provided that the card drawn is actually red, then it seems natural to observe that h_1, but not h_2, receives support, and is thereby confirmed. By (SC), on the contrary, confirmation must extend to any consequence of h_1, including h_2, so that here $\vdash_P^A$ lets in too much. This illustrates a much more general concern. In fact, for *any* pair of unrelated (independent) statements x,y such that the latter is likely enough in its own terms, we will have $x \vdash_P^A (x \wedge y)$ and thus, by (SC), $x \vdash_P^A y$ too.

Let us now turn to (CC). This states that evidence e can never confirm incompatible hypotheses. But consider, by way of illustration, a clinical case of an infectious disease of unknown origin, and suppose that e is the failure of antibiotic treatment. There seems to be nothing wrong in saying that, by discrediting bacteria as possible causes, the evidence confirms (viz. provides support for) any of a number of alternative viral diagnoses. This would not be allowed by (CC), however; so that here $\vdash_P^A$ lets in too little.

In contrast to the foregoing, the following distinctive principles of *incremental* confirmation show that this notion matches widespread patterns of reasoning about evidence and hypotheses.

(CE) *Converse entailment condition*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, if $h \models e$, then $e \mathrel{\sim\!\!\!\vdash}^{I}_{P} h$.

(EComp) *Complementary evidence*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, if $e \mathrel{\sim\!\!\!\vdash}^{I}_{P} h$ then $\neg e \not\mathrel{\sim\!\!\!\vdash}^{I}_{P} h$.

Condition (CE) (the label once again comes from Hempel 1945: p. 104) naturally conveys the statement that hypotheses are confirmed by their consequences that are borne out by observation, this being a paramount precept of scientific methodology. Notably, this elementary principle would not be licenced by the relation of absolute confirmation $\mathrel{\sim\!\!\!\vdash}^{A}_{P}$.

Condition (EComp) is no less relevant. Consider the following example. A father is suspected of abusing his child. Suppose that the child does indeed claim that s/he has been abused (label this piece of evidence e). A forensic psychiatrist, when consulted, declares that this confirms guilt (h). Alternatively, suppose that the child is asked and does *not* report having been abused ($\neg e$). As pointed out by Dawes (2001), it may well happen that a forensic psychiatrist will nonetheless interpret *this* as evidence confirming guilt (suggesting that violence has prompted the child's denial). One might want to argue that this kind of "heads I win, tails you lose" judgment would be inconsistent, and thus untenable on a purely logical basis. Whoever concurs with this line of argument (as Dawes 2001 himself did) must be presupposing the incremental, not the absolute, notion of confirmation. The latter would not do, in fact, for it is easy to show that $e \mathrel{\sim\!\!\!\vdash}^{A}_{P} h$ and $\neg e \mathrel{\sim\!\!\!\vdash}^{A}_{P} h$ can obtain concurrently. Condition (EComp), on the other hand, prescribes that *only one* of the contradictory statements e and $\neg e$ can (incrementally) confirm a hypothesis h.

Remarks such as the foregoing have induced some contemporary theorists to dismiss the very notion of absolute confirmation, concluding that "if you had $P(h|e)$ close to unity [i.e., $e \mathrel{\sim\!\!\!\vdash}^{A}_{P} h$, in our current notation], but less than $P(h)$ [i.e., $e \not\mathrel{\sim\!\!\!\vdash}^{I}_{P} h$], you *ought not* to say that h was confirmed by e" (Good 1968: p. 134; see also Salmon 1975: p. 13). In the remainder of this chapter, we will comply with this suggestion and focus on confirmation in the incremental sense throughout.

30.2 The Axiomatics of Quantitative Confirmation

Assessments of the *amount* of support that a piece of evidence brings to a hypothesis are commonly required in scientific reasoning, as well as in other domains, if only in the form of comparative judgments such as "hypothesis h is more strongly confirmed by e_1 than by e_2" or "e confirms h_1 to a greater extent than h_2." A purely qualitative theory of confirmation is not up to the challenge of providing a foundation for judgments of this kind. However, a probabilistic approach does allow for a proper quantitative treatment, i.e., the definition of

a *measure* $C_P(h,e)$: $\{L_c \times L_c \times \boldsymbol{P}\} \rightarrow \Re$ of the degree of confirmation that h receives from e relative to P. (Indeed, as we shall see shortly, a wealth of such measures can be proposed.)

As we want a confirmation measure $C_P(h,e)$ to have relevant probabilities as its building blocks, the following background assumption is in order:

(F) *Formality*
There exists a function g such that, for any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e) = g[P(h\wedge e),P(h),P(e)]$.

Note that the probability distribution over the algebra generated by h and e is entirely determined by $P(h\wedge e)$, $P(h)$, and $P(e)$. Hence (F) simply states that $C_P(h,e)$ depends on that distribution, and nothing else. This is a widespread assumption in discussions of confirmation in a probabilistic framework, although it is often tacit or spelled out in slightly different ways. (The label *formality* is taken from Tentori, Crupi, and Osherson 2007, 2010).

Another preliminary constraint is sometimes defined along the following lines:

(D) *Discrimination*
There exists $t \in \Re$ such that, for any $h,e \in L_c$ and any $P \in \boldsymbol{P}$:

(i) $C_P(h,e) > t$ if and only if $e \mathrel{\overset{I}{P}\!\!\sim} h$;
(ii) $C_P(h,e) < t$ if and only if $e \mathrel{\overset{I}{P}\!\!\sim} \neg h$;
(iii) $C_P(h,e) = t$ if and only if $e \not\mathrel{\overset{I}{P}\!\!\sim}$ and $e \not\mathrel{\overset{I}{P}\!\!\sim} \neg h$.

Principle (D) states that a fixed figure t acts as a threshold separating cases in which e confirms h (thus *dis*confirming $\neg h$, as we will say hereafter) from cases in which e confirms $\neg h$ (thus *dis*confirming h). The value t itself indicates *neutrality* (of evidence e relative to h vs. $\neg h$) and is set as a matter of convenience, usual choices being 0 or 1. Condition (D) suffices to guarantee that the foregoing properties of $\mathrel{\overset{I}{P}\!\!\sim}\!-$ as conveyed by (EC), (NM), (Cases), (CComp), (CE), and (EComp) above – are all retained under $C_P(h,e)$, thus fulfilling a natural constraint of coherence between the purely qualitative notion and its quantitative refinement. (D) is a rather mild requirement, however, for there exist functions of all sorts that satisfy it. Historically, the outlook of theorists for the representation of $C_P(h,e)$ has been much more selective. The most popular candidates have in fact amounted to the following:[5]

Probability difference: $P(h|e) - P(h)$
Probability ratio: $P(h|e)/P(h)$
Likelihood ratio: $P(e|h)/P(e|\neg h)$

Although they are all consistent with (D), the above quantities differ substantially in that they are *not ordinally equivalent*. Two confirmation measures are said to be ordinally equivalent if they *always rank* evidence–hypothesis pairs in the same way. More formally, $C_P(h,e)$ and $C^*_P(h,e)$ are ordinally equivalent if and only if, for any $h_1,h_2,e_1,e_2 \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h_1,e_1) \gtreqless C_P(h_2,e_2)$ if and only if $C^*_P(h_1,e_1) \gtreqless C^*_P(h_2,e_2)$. Isotone transformations of a given quantity yield measures whose detailed quantitative behavior (including range

[5] The probability difference has been first defined by Carnap (1950/1962: p. 361), the probability ratio by Keynes (1921: pp. 165 ff.), and the likelihood ratio by Alan Turing (as reported in Good 1950: pp. 62–3).

and neutrality value) may vary widely, but such that rank-order is strictly preserved. For instance, the measures in the following list are all ordinally equivalent variants based on the probability ratio:[6]

$r_0(h,e) = P(h\|e)/P(h)$	range: $[0,+\infty)$	neutrality value: 1
$r_1(h,e) = \frac{P(h\|e)-P(h)}{P(h)}$	range: $[-1,+\infty)$	neutrality value: 0
$r_2(h,e) = \ln[P(h\|e)/P(h)]$	range: $[-\infty,+\infty)$	neutrality value: 0
$r_3(h,e) = \frac{P(h\|e)-P(h)}{P(h\|e)+P(h)}$	range: $[-1,1)$	neutrality value: 0
$r_4(h,e) = \frac{P(h\|e)}{P(h\|e)+P(h)}$	range: $[0,1)$	neutrality value: ½

The ordinal divergence among alternative confirmation measures is arguably of greater theoretical significance than purely quantitative differences, because the former, unlike the latter, implies opposite comparative judgments for some evidence-hypothesis pairs. Indeed, in what follows we will deal *only* with the ordinal level in the assessment of confirmation. We will thus invariably address properties that apply to $C_P(h,e)$ if and only if they also apply to any ordinally equivalent $C^*_P(h,e)$ and treat classes of ordinal equivalence as our unit of analysis. Accordingly, we will simply say that $C_P(h,e)$ is a *probability difference measure* if and only if there exists a strictly increasing function f such that $C_P(h,e) = f[P(h|e) - P(h)]$, and the same for the probability and likelihood ratio.

An effective tool to gain theoretical insight concerning (ordinally) different measures of confirmation is to provide an exhaustive set of axioms for each of them. It turns out that four fundamental statements, along with the basic requirement of formality (F), suffice to distinguish neatly among the traditional options considered above.[7]

(C1) *Final probability*
For any $h,e_1,e_2 \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e_1) \gtreqless C_P(h,e_2)$ if and only if $P(h|e_1) \gtreqless P(h|e_2)$.

(C2) *Disjunction of alternative hypotheses*
For any $h_1,h_2,e \in L_c$ and any $P \in \boldsymbol{P}$, if $P(h_1 \wedge h_2) = 0$, then $C_P(h_1 \vee h_2,e) \gtreqless C_P(h_1,e)$ if and only if $P(h_2|e) \gtreqless P(h_2)$.

(C3) *Law of likelihood*
For any $h_1,h_2,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h_1,e) \gtreqless C_P(h_2,e)$ if and only if $P(e|h_1) \gtreqless P(e|h_2)$.

(C4) *Modularity (for conditionally independent data)*
For any $h,e_1,e_2 \in L_c$ and any $P \in \boldsymbol{P}$, if $P(e_1|\pm h \wedge e_2) = P(e_1|\pm h)$, then $C_P(h,e_1|e_2) = C_P(h,e_1)$.[8]

(C1) states that, for any hypothesis h, final probability and confirmation always move in the same direction in the light of data, e. This seems a very compelling principle, and it is

[6] Obviously, $r_1(h,e) = r_0(h,e) - 1$ and $r_2(h,e) = \ln[r_0(h,e)]$. Moreover, $r_3(h,e) = [r_0(h,e) - 1]/[r_0(h,e) + 1]$ and $r_4(h,e) = r_0(h,e)/[r_0(h,e) + 1]$.

[7] For recent occurrences of (C1), see Fitelson (2006: p. 506) and Hájek and Joyce (2008: p. 122). (C3) is endorsed by both Edwards (1972: pp. 30–1) and Milne (1996). The label "law of likelihood" goes back to Hacking (1965), while that for (C4) is freely adapted from Heckerman (1988: pp. 18–19).

[8] The notion of conditional confirmation denoted by $C_P(h,e_1|e_2)$ implies that all relevant values from P are conditionalized on e_2. The expression "$\pm h$" is meant to cover two cases, i.e., both the statement and the negation of h. In some contexts, condition $P(e_1|\pm h \wedge e_2) = P(e_1|\pm h)$ is also referred to as *screening off* (of e_1 and e_2 by h).

in fact the only condition among the foregoing that has remained virtually unchallenged.[9] On the other hand, the choice among the competing measures listed essentially depends on the acceptance of either (C2), (C3), or (C4), as shown by the result below. (Note that one does not need to assume the fundamental *Discrimination* condition (D) separately, for it follows in any of the three clauses of the theorem and thus becomes formally redundant.)

Theorem 1

(i) *(F), (C1) and (C2) if and only if $C_P(h,e)$ is a probability difference measure.*
(ii) *(F), (C1) and (C3) if and only if $C_P(h,e)$ is a probability ratio measure.*
(iii) *(F), (C1) and (C4) if and only if $C_P(h,e)$ is a likelihood ratio measure.*

A proof of clause (i) of Theorem 1 is given in the Appendix, while clauses (ii) and (iii) are proven in Crupi, Chater, and Tentori (2013).[10]

The plurality of probabilistic measures of confirmation has prompted some scholars to be skeptical or dismissive of the prospects for a quantitative theory of confirmation (see, e.g., Howson 2000: pp. 184–5, and Kyburg and Teng 2001: pp. 98 ff.). However, quantitative probabilistic analyses have proved crucial for handling a number of puzzles and issues that plagued more qualitative approaches, including the so-called "irrelevant conjunction" problem, Hempel's paradox of the ravens, Goodman's "new riddle of induction", the variety of evidence, and the Duhem-Quine thesis (see Earman 1992: pp. 63–117 for a now classic discussion, and Crupi 2015 for a more recent survey). And in fact, various arguments in the philosophy of science have been shown to depend critically (and sometimes unwittingly) on the choice of one confirmation measure (or some of them) rather than others (Festa 1999; Fitelson 1999; Brössel 2013). Relying on the appeal of distinctive features, some authors have insisted on "one true measure" of confirmation (see Good 1984; Milne 1996; but also see Milne 2012), while others have seen different measures as possibly capturing "distinct,

[9] Precisely for this reason, we forgo detailed treatment of candidate measures departing from (C1). To be noted, however, that among these Carnap's (1950/1962: p. 360) measure $P(h\wedge e) - P(h)P(e)$ implies (C2), Mortimer's (1988: Section 11.1) measure $P(e|h) - P(e)$ implies (C3), and Nozick's (1981: p. 252) measure $P(e|h) - P(e|\neg h)$ implies (C4). Indeed, the corresponding classes of ordinally equivalent measures can be axiomatized much as in our Theorem 1, provided that (C1) is replaced with the following (proofs omitted):

(C1*) Disjunction of alternative data
For any $h,e_1,e_2 \gtreqless \in L_c$ and any $P \in \boldsymbol{P}$, if $P(e_1\wedge e_2) = 0$, then $C_P(h|e_1\vee e_2) \gtreqless C_P(h,e_1)$ if and only if $P(h|e_2) \gtreqless P(h)$.

[10] In line with Carnap's (1950/1962) classic work, the standard quantitative counterpart of *absolute* qualitative confirmation (namely, relation $\vdash_P^A$ from the preceding section) is $P(h|e)$ itself. One can thus wonder whether also this notion is amenable to a similar axiomatic treatment. To see that this is in fact the case, consider the following condition (already put forward by Törnebohm 1966: p. 81):

(A) For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e) = C_P(h\wedge e,e)$.

It can then be shown that (F), (C1) and (A) hold if and only if there exists a strictly increasing function f such that $C_P(h,e) = f[P(h|e)]$ (see Schippers 2016). Condition (A) seems indeed plausible *if* (but *only if*, in our view) the overall credibility of the hypothesis on the evidence is at issue (as contrasted with the *impact* on the credibility of the hypothesis yielded by the evidence). Accordingly, (A) is inconsistent with (D) above.

complementary notions of evidential support" (Hájek and Joyce 2008: p. 123; also see Huber 2008).

We find the latter approach sensible, but suggest that pluralism be supplemented and tempered by critical scrutiny (see Steel 2007 for a similar position). The axiomatic characterization of competing measures seems particularly useful for this purpose. By way of illustration, once again consider a draw from a standard well shuffled deck and posit h_1 = "the card drawn is 7♠", h_2 = "the card drawn is red", and e = "the card drawn is a face". Note that here $P(h_2|e) = P(h_2)$, while the antecedent of (C2) is satisfied, so according to this principle $C_P(h_1,e) = C_P(h_1 \vee h_2,e)$, even if h_1 is conclusively disconfirmed (i.e., plainly refuted) by e, while $h_1 \vee h_2$ is not (for any red face card would still make it true notwithstanding e). This implication might well seem disturbing, thus speaking against difference measures of confirmation. For, as pointed out by Zalabardo (2009: p. 631), a "choice between […] accounts of confirmation should be dictated by the plausibility of the orderings they generate". For a (critical) discussion of (C3) and a (supportive) examination of (C4), both carried out in a similar vein, we refer the reader to Fitelson (2007 and 2001, respectively).

30.3 Confirmation as Partial Entailment Relative Distance Measures

It has often been maintained that confirmation theory should yield an *inductive logic* that is analogous to classical deductive logic in some suitable sense. This view has been pursued in a number of variants, mostly depending, as Hawthorne (2014) has observed, on "precisely how the deductive model is emulated". According to an old and illustrious idea, the deductive model should be paralleled by a generalized, quantitative theory of *partial entailment*. The following revealing passage, again from Hawthorne (2014), attests to the enduring influence of this notion, albeit from a pessimistic perspective:

> A collection of premise sentences *logically entails* a conclusion sentence just when the negation of the conclusion is *logically inconsistent* with those premises. An inductive logic must, it seems, deviate from this paradigm […]. Although the notion of *inductive support* is analogous to the deductive notion of *logical entailment*, and is arguably an extension of it, there seems to be no inductive logic extension of the notion of *logical inconsistency* – at least none that is interdefinable with *inductive support* in the way that *logical inconsistency* is interdefinable with *logical entailment*.

The central point of the present section amounts to showing that this resignation is overly hasty. As we shall see, it is perfectly possible to have a sound extension of the notion of logical inconsistency that is indeed interdefinable with inductive support in essentially the same way that logical inconsistency is interdefinable with logical entailment. So much so, in fact, that one can safely and fruitfully embed into axioms those very properties which inductive logic would inevitably lack, according to Hawthorne's line of argument. To show this, we first need to introduce a new class of confirmation measures. Consider the following function:

$$z(h,e) = \begin{cases} \frac{P(h|e)-P(h)}{1-P(h)} & \text{if} \quad P(h|e) \geq P(h) \\ \frac{P(h|e)-P(h)}{P(h)} & \text{if} \quad P(h|e) < P(h) \end{cases}$$

Despite its twofold algebraic form, measure $z(h,e)$ conveys a unifying core intuition.[11] More precisely, in case of (positive) confirmation, $z(h,e)$ measures how far upward the posterior $P(h|e)$ has gone in covering the distance between the prior $P(h)$ and 1; that is, it expresses the relative reduction of the initial distance from certainty of h being true as yielded by e. Similarly, in the case of *dis*confirmation, $z(h,e)$ measures how far downward the posterior $P(h|e)$ has gone in covering the distance between the prior $P(h)$ and 0; that is, it reflects the relative reduction of the initial distance from certainty of h being false as yielded by e. So $z(h,e)$ measures the extent to which the initial probability distance from certainty concerning the truth/falsehood of h is reduced by the confirming/disconfirming statement e. Or, put otherwise, how much of such distance is "covered" by the upward/downward jump from $P(h)$ to $P(h|e)$. $z(h,e)$ is thus a measure of the relative reduction of the distance from certainty that a hypothesis of interest is true or false – or, in short (and abusing language a little), a *relative distance* measure.[12] Accordingly, we will say that $C_P(h,e)$ is a *relative distance measure* if and only if there exists a strictly increasing function f such that $C_P(h,e) = f[z(h,e)]$.[13]

Relying again on an axiomatic approach, we now show how relative distance measures escape Hawthorne's (2014) pessimistic conclusion. Drawing from the quote above, we will first assume $C_P(h,e)$ to exhibit a commutative behavior whenever h and e are inductively at odds (i.e., negatively correlated), thus paralleling the symmetric nature of logical inconsistency, as follows:

(C5) *Partial inconsistency*
For any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, if $P(h \wedge e) \leq P(h)P(e)$, then $C_P(h,e) = C_P(e,h)$.

An unrestricted form of commutativity has appeared as a basic and sound requirement in probabilistic analyses of degrees of "coherence" (and lack thereof).[14] In (C5), however, commutativity is not meant to extend to the quantification of *positive* confirmation or support, because logical entailment (unlike refutation) is not symmetric; nor is it coextensive with logical equivalence (or mere logical consistency, for that matter) in the way that refutation is coextensive with inconsistency (Eells and Fitelson 2002 and Crupi, Tentori, and Gonzalez 2007 discuss this point further).

As for the interdefinability of logical entailment of h from e and inconsistency of e with $\neg h$, it naturally generalizes to an inverse (ordinal) correlation between positive confirmation and partial inconsistency with regard to complementary hypotheses, as follows:

(C6) *Confirmation complementarity* (*ordinal*)
For any $h_1,h_2,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h_1,e) \gtreqless C_P(h_2,e)$ if and only if $C_P(\neg h_1,e) \lesseqgtr C_P(\neg h_2,e)$.

[11] An alternative, more compact rendition is the following:

$$z(h,e) = \frac{\min[P(h|e),P(h)]}{P(h)} - \frac{\min[P(\neg h|e),P(\neg h)]}{P(\neg h)}$$

In this form, $z(h,e)$ is structurally similar to Mura's (2006, 2008) measure of "partial entailment". Mura's measure and $z(h,e)$, however, are demonstrably non-equivalent in ordinal terms (see Crupi and Tentori 2014).

[12] This label was first adopted by Huber (2007).

[13] For more extensive discussion and some additional relevant references, see Crupi, Tentori, and Gonzalez (2007), Crupi, Festa, and Buttasi (2010), and Crupi and Tentori (2010, 2013).

[14] See Shogenji's (1999) seminal work. For updated and informed discussions of subsequent developments, see Schupbach (2011) and Schippers (2014).

Indeed, (C6) can be viewed as a fairly faithful formal rendition of Keynes' (1921: p. 80) remark that "an argument is always as near to proving or disproving a proposition, as it is to disproving or proving its contradictory".

To sum up, (C5) implies that, when h and e are at odds, lower values of confirmation measure $C_P(h,e)$ precisely amount to a higher degree of partial mutual inconsistency. (C6), on the other hand, implies that the positive confirmation or support from e to h is in fact nothing other than a strictly increasing function of the degree of partial inconsistency between e and $\neg h$. Hawthorne's (2014) aforesaid "impossibility" statement would suggest that no sensible probabilistic analysis of confirmation could satisfy such requirements, no matter how appealing they may seem. The following result, however, opens up a different scenario (see Crupi and Tentori 2013 for a proof):

Theorem 2 *(F), (C1), (C5) and (C6) if and only if $C_P(h,e)$ is a relative distance measure.*

As pointed out earlier, probabilistic measures of incremental confirmation are known to be many and diverse. Whatever the amount of pluralism that one is willing to allow for in this respect, Theorem 2 shows that a small set of properties single out relative distance measures as uniquely capturing the notion of partial entailment (and refutation). In fact, whilst all the alternatives discussed above satisfy the fundamental assumptions (F) and (C1), they demonstrably fail to capture either (C5) or (C6), thus falling within the scope of Hawthorne's (2014) negative conclusion. And yet this conclusion does not hold unrestrictedly – a genuine confirmation-theoretic generalization of logical entailment (and refutation) is possible after all.

Acknowledgements

Research relevant to this work has been supported by the Deutsche Forschungsgemeinshaft (DFG) as part of the prority program *New Frameworks of Rationality* (SPP 1516, Grant CR 409/1-1), and by the Italian Ministry of Scientific Research (FIRB project *Structures and Dynamics of Knowledge and Cognition*, Turin unit, D11J12000470001, and PRIN grant 2010RP5RNM_006). We thank Chris Hitchcock and Gustavo Cevolani for very useful comments on previous drafts of this chapter.

References

Antonelli, G. A. (2012) Non-monotonic Logic. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/archives/win2012/entries/logic-nonmonotonic/. [Accessed 23 Sep 2015.]

Brössel, P. (2013) The Problem of Measure Sensitivity Redux. *Philosophy of Science*. 80. pp. 378–97.

Carnap, R. (1950/1962) *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Crupi, V. (2015) Confirmation. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/archives/fall2015/entries/confirmation/. [Accessed 23 Sep 2015.]

Crupi, V., Chater, N., and Tentori, K. (2013) New Axioms for Probability and Likelihood Ratio Measures. *British Journal for the Philosophy of Science*. 64. pp. 189–204.

Crupi, V., Festa, R., and Buttasi, C. (2010) Towards a Grammar of Bayesian Confirmation. In Suárez, M., Dorato, M., and Rédei, M. (eds.) *Epistemology and Methodology of Science*. pp. 73–93. Dordrecht: Springer.

Crupi, V., Fitelson, B., and Tentori, K. (2008) Probability, Confirmation, and the Conjunction Fallacy. *Thinking and Reasoning*. 14. pp. 182–99.

Crupi, V. and Tentori, K. (2010) Irrelevant Conjunction: Statement and Solution of a New Paradox. *Philosophy of Science*. 77. pp. 1–13.

Crupi, V. and Tentori, K. (2013) Confirmation as Partial Entailment: A Representation Theorem in Inductive Logic. *Journal of Applied Logic*. 11. pp. 364–72. [Erratum in *Journal of Applied Logic*. 12. (2014). pp. 230–1].

Crupi, V. and Tentori, K. (2014) Measuring information and confirmation. *Studies in the History and Philosophy of Science*. 47. pp. 81–90.

Crupi, V., Tentori, K., and Gonzalez, M. (2007) On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues. *Philosophy of Science*. 74. pp. 229–52.

Dawes, R. M. (2001) *Everyday Irrationality*. Boulder, CO: Westview.

Earman, J. (1992) *Bayes or bust?*, Cambridge, MA: MIT Press.

Edwards, A. W. F. (1972) *Likelihood*. Cambridge: Cambridge University Press.

Eells, E. and Fitelson, B. (2002) Symmetries and Asymmetries in Evidential Support. *Philosophical Studies*. 107. pp. 129–42.

Festa, R. (1999) Bayesian Confirmation. In Galavotti, M. and Pagnini, A. (eds.) *Experience, Reality, and Scientific Explanation*. pp. 55–87. Dordrecht: Kluwer.

Fitelson, B. (1999) The Plurality of Bayesian Measures of Confirmation and the Problem of Measure Sensitivity. *Philosophy of Science*. 66. S362–78.

Fitelson, B. (2001) A Bayesian Account of Independent Evidence with Applications. *Philosophy of Science*. 68. S123–40.

Fitelson, B. (2006) Logical Foundations of Evidential Support. *Philosophy of Science*. 73. pp. 500–12.

Fitelson, B. (2007) Likelihoodism, Bayesianism, and Relational Confirmation. *Synthese*. 156. pp. 473–89.

Fitelson, B. and Hawthorne, J. (2010) The Wason Task(s) and the Paradox of Confirmation. *Philosophical Perspectives*. 24. pp. 207–41.

Gabbay, D. M. (1982) Intuitionistic Basis for Non-monotonic Logic. In Goos, G. and Hartmanis, J. (eds.) *Proceedings of the 6th Conference on Automated Deduction*, *Lecture Notes in Computer Science*. 138. pp. 260–73.

Good, I. J. (1950) *Probability and the Weighing of Evidence*. London: Griffin.

Good, I. J. (1968) Corroboration, Explanation, Evolving Probabilities, Simplicity, and a Sharpened Razor. *British Journal for the Philosophy of Science*. 19. pp. 123–43.

Good, I. J. (1984) The Best Explicatum for Weight of Evidence. *Journal of Statistical Computation and Simulation*. 19. pp. 294–9.

Hacking, I. (1965) *Logic of Statistical Inference*. Cambridge, MA: Cambridge University Press.

Hájek, A. and Joyce, J. (2008) Confirmation. In Psillos, S. and Curd, M. (eds.) *Routledge Companion to the Philosophy of Science*. pp. 115–29. New York, NY: Routledge.

Hawthorne, J. (2014) Inductive Logic. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/archives/sum2014/entries/logic-inductive/. [Accessed 23 Sep 2015.]

Heckerman, D. (1988) An Axiomatic Framework for Belief Updates. In Lemmer, J. F. and Kanal, L. N. (eds.) *Uncertainty in Artificial Intelligence*. 2. pp. 11–22. Amsterdam: North-Holland.

Hempel, C. G. (1943) A Purely Syntactical Definition of Confirmation. *Journal of Symbolic Logic*. 8. pp. 122–43.

Hempel, C. G. (1945) Studies in the Logic of Confirmation. *Mind*. 54. pp. 1–26, 97–121.

Howson, C. (2000) *Hume's Problem: Induction and the Justification of Belief*. New York, NY: Oxford University Press.

Huber, F. (2007) Confirmation and Induction. In *Internet Encyclopedia of Philosophy*. [Online] Available from: http://www.iep.utm.edu/conf-ind/#SH6b. [Accessed 23 Sep 2015.]

Huber, F. (2008) Milne's Argument for the Log-Ratio Measure. *Philosophy of Science*. 75. pp. 413–20.

Kelly, K. T. and Glymour, C. (2004) Why Probability Does not Capture the Logic of Scientific Justification. In Hitchcock, C. (ed.) *Contemporary Debates in the Philosophy of Science*. pp. 94–114. London: Blackwell.

Keynes, J. (1921) *A Treatise on Probability*. London: Macmillan.

Kuipers, T. (2000) *From Instrumentalism to Constructive Realism*. Dordrecht: Reidel.

Kuipers, T. (2007) The Hypothetico-Probabilistic (HP-) Method as a Concretization of the HD-Method. In Sintonen, M., Pihlström, S., and Raatikainen, P. (eds.) *Festschrift in Honor of Ilkka Niiniluoto*. pp. 179–207. London: King's College .

Kyburg, H. E. and Teng, C. M. (2001) *Uncertain Inference*. New York, NY: Cambridge University Press.

Levi, I. (2010) Probability Logic, Logical Probability, and Inductive Support. *Synthese*. 172. pp. 97–118.

Malinowski, J. (2005) Logic of Simpson Paradox. *Logic and Logical Philosophy*. 14. pp. 203–10.

Mayo, D. (1996) *Error and the Growth of Experimental Knowledge*. Chicago, IL: University of Chicago Press.

Milne, P. (1996) Log[$P(h|eb)/P(h|b)$] Is the One True Measure of Confirmation. *Philosophy of Science*. 63. pp. 21–6.

Milne, P. (2012) (ms.) On Measures of Confirmation. Unpublished manuscript.

Mortimer, H. (1988) *The Logic of Induction*. Paramus, NJ: Prentice Hall.

Mura, A. (2006) Deductive Probability, Physical Probability, and Partial Entailment. In Alai, M. and Tarozzi, G. (eds.) *Karl Popper: Philosopher of Science*. pp. 181–202. Soveria Mannelli: Rubbettino.

Mura, A. (2008) Can Logical Probability Be Viewed as a Measure of Degrees of Partial Entailment? *Logic & Philosophy of Science*. 6. pp. 25–33.

Norton, J. D. (2010) There Are no Universal Rules for Induction. *Philosophy of Science*. 77. pp. 765–77.

Nozick, R. (1981) *Philosophical Explanations*. Oxford: Clarendon Press.

Popper, K. R. (1959) *The Logic of Scientific Discovery*. London: Routledge.

Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Salmon, W. C. (1969) Partial Entailment as a Basis for Inductive Logic. In Rescher, N. (ed.) *Essays in Honor of Carl G. Hempel*. pp. 47–81. Dordrecht: Reidel.

Salmon, W. C. (1975) Confirmation and Relevance. In Maxwell, G. and Anderson, R. M. Jr. (eds.) *Induction, Probability, and Confirmation: Minnesota Studies in the Philosophy of Science*. 6. pp. 3–36. Minneapolis, MN: University of Minnesota Press.
Schippers, M. (2014) Probabilistic Measures of Coherence: From Adequacy Constraints Towards Pluralism. *Synthese*. 191. pp. 3821–45
Schippers, M. (2016) A Representation Theorem for Absolute Confirmation. *Philosophy of Science*. Forthcoming.
Schupbach, J. (2011) New Hope for Shogenji's Coherence Measure. *British Journal for the Philosophy of Science*. 62. pp. 125–42.
Shogenji, T. (1999) Is Coherence Truth Conducive? *Analysis*. 59. pp. 338–45.
Steel, D. (2007) Bayesian Confirmation Theory and the Likelihood Principle. *Synthese*. 156. pp. 55–77.
Tentori, K., Crupi, V., and Osherson, D. (2007) Determinants of Confirmation. *Psychonomic Bulletin & Review*. 14. pp. 877–83.
Tentori, K., Crupi, V., and Osherson, D. (2010) Second-order Probability Affects Hypothesis Confirmation. *Psychonomic Bulletin & Review*. 17. pp. 129–34.
Törnebohm, H. (1966) Two Measures of Evidential Strength. In Hintikka, J. and Suppes, P. (eds.) *Aspects of Inductive Logic*. pp. 81–95. Amsterdam: North-Holland.
Zalabardo, J. (2009) An Argument for the Likelihood Ratio Measure of Confirmation. *Analysis*. 69. pp. 630–5.

APPENDIX

Theorem 1. Clause (i).

(F), (C1) and (C2) if and only if $C_P(h,e)$ is a probability difference measure.

Proof

The proof provided concerns the left-to-right implication (verification of the right-to-left implication is simple).

Notice that $P(h\wedge e) = \{[P(h|e) - P(h)] + P(h)\}P(e)$. As a consequence, by (F), there exists a function j such that, for any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e) = j[P(h|e) - P(h),P(h),P(e)]$. With no loss of generality, we will enforce probabilistic coherence and regularity by constraining the domain of j to include triplets of values (x,y,w) such that the following conditions are jointly satisfied:

- $0 < y,w < 1$;
- $x \geq -y$, by which $x + y = P(h|e) \geq 0$, and thus $P(h\wedge e) \geq 0$;
- $x \leq 1 - y$, by which $x + y = P(h|e) \leq 1$, so that $P(h\wedge e) \leq P(e)$, and thus $P(\neg h\wedge e) \geq 0$;
- $x \leq y(1/w - 1)$, by which $[(x + y)w]/y = P(e|h) \leq 1$ so that $P(h\wedge e) \leq P(h)$, and thus $P(h\wedge\neg e) \geq 0$;
- $x \geq (1 - y)(1 - 1/w)$, by which $(x + y)w = P(h\wedge e) \geq P(h) + P(e) - 1 = y + w - 1$, and thus $P(h\wedge e) + P(\neg h\wedge e) + P(h\wedge\neg e) \leq 1$.

We thus posit j: $\{(x,y,w) \in \{(-1,+1) \times (0,1)^2 \mid -y, (1 - y)(1 - 1/w) \leq x \leq y(1/w - 1), 1 - y\} \rightarrow \Re$ and denote the domain of j as D_j.

Lemma 1. For any x,y,w_1,w_2 such that $x \in (-1,+1)$, $y,w_1,w_2 \in (0,1)$, and $-y$, $(1-y)(1-1/w_1)$, $(1-y)(1-1/w_2) \leq x \leq y(1/w_1-1)$, $y(1/w_2-1)$, $1-y$, there exist $h,e_1,e_2 \in L_c$ and $P' \in \boldsymbol{P}$ such that $P'(h|e_1) - P'(h) = P'(h|e_2) - P'(h) = x$, $P'(h) = y$, $P'(e_1) = w_1$, and $P'(e_2) = w_2$.

Proof. The equalities in Lemma 1 arise from the following scheme of probability assignments:

$$P'(h \wedge e_1 \wedge e_2) = \frac{(x+y)^2 w_1 w_2}{y};$$

$$P'(h \wedge e_1 \wedge \neg e_2) = (x+y)w_1\left[1 - \frac{(x+y)w_2}{y}\right];$$

$$P'(h \wedge \neg e_1 \wedge e_2) = \left[1 - \frac{(x+y)w_1}{y}\right](x+y)w_2;$$

$$P'(h \wedge \neg e_1 \wedge \neg e_2) = \left[1 - \frac{(x+y)w_1}{y}\right]\left[1 - \frac{(x+y)w_2}{y}\right]y;$$

$$P'(\neg h \wedge e_1 \wedge e_2) = \frac{\left(1-x-y\right)^2 w_1 w_2}{(1-y)};$$

$$P'(\neg h \wedge e_1 \wedge \neg e_2) = \left(1-x-y\right)w_1\left[1 - \frac{\left(1-x-y\right)w_2}{(1-y)}\right];$$

$$P'(\neg h \wedge \neg e_1 \wedge e_2) = \left[1 - \frac{\left(1-x-y\right)w_1}{(1-y)}\right]\left(1-x-y\right)w_2;$$

$$P'(\neg h \wedge \neg e_1 \wedge \neg e_2) = \left[1 - \frac{\left(1-x-y\right)w_1}{(1-y)}\right]\left[1 - \frac{\left(1-x-y\right)w_2}{(1-y)}\right](1-y).$$

Suppose there exist (x,y,w_1), $(x,y,w_2) \in D_j$ such that $j(x,y,w_1) \neq j(x,y,w_2)$. Then, by Lemma 1 and the definition of D_j, there exist $h,e_1,e_2 \in L_c$ and $P' \in \boldsymbol{P}$ such that $P'(h|e_1) - P'(h) = P'(h|e_2) - P'(h) = x$, $P'(h) = y$, $P'(e_1) = w_1$ and $P'(e_2) = w_2$. Clearly, if the latter equalities hold, then $P'(h|e_1) = P'(h|e_2)$. Thus, there exist $h,e_1,e_2 \in L_c$ and $P' \in \boldsymbol{P}$ such that $C_{P'}(h,e_1) = j(x,y,w_1) \neq j(x,y,w_2) = C_{P'}(h,e_2)$ even if $P'(h|e_1) = P'(h|e_2)$, contradicting (C1). Conversely, (C1) implies that, for any (x,y,w_1), $(x,y,w_2) \in D_j$, $j(x,y,w_1) = j(x,y,w_2)$. So, for (C1) to hold, there must exist k such that, for any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e) = k[P(h|e) - P(h),P(h)]$ and $k(x,y) = j(x,y,w)$. We thus posit k: $\{(x,y) \in \{(-1,+1) \times (0,1) \mid -y \leq x \leq 1-y\} \rightarrow \Re$ and denote the domain of k as D_k.

Lemma 2. For any x,y_1,y_2 such that $x \in (-1,+1)$, $y_1,y_2 \in (0,1)$, $-y_1 \leq x \leq 1-y_2$, and $y_1 < y_2$, there exist $h_1,h_2,e \in L_c$ and $P'' \in \boldsymbol{P}$ such that $P''(h_1|e) - P''(h_1) = x$, $P''(h_1) = y_1$, $P''(h_1 \vee h_2) = y_2$, $P''(h_2|e) = P''(h_2)$, and $P''(h_1 \wedge h_2) = 0$.

Proof. Let $w \in (0,1)$ be given so that $w \leq y_1/(x+y_1)$, $(1-y_2)/(1-x-y_2)$ (as the latter quantities must be positive, w exists), and posit $h_2 = \neg h_1 \wedge q$, with q an atomic sentence in L_c and $q \neq h_1$. The equalities in Lemma 2 arise from the following scheme of probability assignments:

$$P''(h_1 \wedge q \wedge e) = (1/2)(x+y_1)w;$$

$$P''(\neg h_1 \wedge q \wedge e) = (y_2 - y_1)w;$$

$$P''(h_1 \wedge q \wedge \neg e) = (1/2)y_1 - (1/2)(x+y_1)w;$$

$$P''(\neg h_1 \wedge q \wedge \neg e) = (y_2 - y_1)(1-w);$$

$$P''(h_1 \wedge \neg q \wedge e) = (1/2)(x + y_1)w;$$
$$P''(\neg h_1 \wedge \neg q \wedge e) = (1 - x - y_2)w;$$
$$P''(h_1 \wedge \neg q \wedge \neg e) = (1/2)y_1 - (1/2)(x + y_1)w;$$
$$P''(\neg h_1 \wedge \neg q \wedge \neg e) = (1 - y_2) - (1 - x - y_2)w.$$

Suppose there exist (x,y_1), $(x,y_2) \in D_k$ such that $k(x,y_1) \neq k(x,y_2)$. Assume $y_1 < y_2$ with no loss of generality. Then, by Lemma 2 and the definition of D_k, there exist $h_1,h_2,e \in L_c$ and $P'' \in \boldsymbol{P}$ such that $P''(h_1|e) - P''(h_1) = x$, $P''(h_1) = y_1$, $P''(h_1 \vee h_2) = y_2$, $P''(h_2|e) = P''(h_2)$, and $P''(h_1 \wedge h_2) = 0$. By the probability calculus, if the latter equalities hold, then $P''(h_1|e) - P''(h_1) = P''(h_1 \vee h_2|e) - P''(h_1 \vee h_2) = x$. Thus, there exist $h_1,h_2,e \in L_c$ and $P'' \in \boldsymbol{P}$ such that $P''(h_1 \wedge h_2) = 0$ and $C_{P''}(h_1,e) = k(x,y_1) \neq k(x,y_2) = C_{P''}(h_1 \vee h_2,e)$ even if $P''(h_2|e) = P''(h_2)$, contradicting (C2). Conversely, (C2) implies that, for any (x,y_1), $(x,y_2) \in D_k$, $k(x,y_1) = k(x,y_2)$. So, for (C2) to hold, there must exist f such that, for any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e) = f[P(h|e) - P(h)]$ and $f(x) = k(x,y)$. We thus posit $f\colon (-1,+1) \to \Re$ and denote the domain of f as D_f.

Lemma 3. For any x_1,x_2 such that $x_1,x_2 \in (-1,+1)$ and $0 \leq x_1 - x_2 < 1$, there exist $h,e_1,e_2 \in L_c$ and $P''' \in \boldsymbol{P}$ such that $P'''(h|e_1) - P'''(h) = x_1$ and $P'''(h|e_2) - P'''(h) = x_2$.

Proof. Let $y,w_1,w_2 \in (0,1)$ be given so that $-x_2 \leq y \leq 1 - x_1$ (as $0 \leq x_1 - x_2 < 1$, y exists), $w_1 \leq y(x_1 + y)$, $(1 - y)/(1 - x_1 - y)$ (as the latter quantities must be positive, w_1 exists), and $w_2 \leq y(x_2 + y)$, $(1 - y)/(1 - x_2 - y)$ (as the latter quantities must be positive, w_2 exists). The equalities in Lemma 3 arise from the following scheme of probability assignments:

$$P'''(h \wedge e_1 \wedge e_2) = \frac{(x_1 + y)(x_2 + y)w_1w_2}{y};$$
$$P'''(h \wedge e_1 \wedge \neg e_2) = (x_1 + y)w_1\left[1 - \frac{(x_2 + y)w_2}{y}\right];$$
$$P'''(h \wedge \neg e_1 \wedge e_2) = \left[1 - \frac{(x_1 + y)w_1}{y}\right](x_2 + y)w_2;$$
$$P'''(h \wedge \neg e_1 \wedge \neg e_2) = \left[1 - \frac{(x_1 + y)w_1}{y}\right]\left[1 - \frac{(x_2 + y)w_2}{y}\right]y;$$
$$P'''(\neg h \wedge e_1 \wedge e_2) = \frac{(1 - x_1 - y)(1 - x_2 - y)w_1w_2}{(1 - y)};$$
$$P'''(\neg h \wedge e_1 \wedge \neg e_2) = (1 - x_1 - y)w_1\left[1 - \frac{(1 - x_2 - y)w_2}{(1 - y)}\right];$$
$$P'''(\neg h \wedge \neg e_1 \wedge e_2) = \left[1 - \frac{(1 - x_1 - y)w_1}{(1 - y)}\right](1 - x_2 - y)w_2;$$
$$P'''(\neg h \wedge \neg e_1 \wedge \neg e_2) = \left[1 - \frac{(1 - x_1 - y)w_1}{(1 - y)}\right]\left[1 - \frac{(1 - x_2 - y)w_2}{(1 - y)}\right](1 - y).$$

Let $x_1,x_2 \in D_f$ be given so that $x_1 - x_2 > 0$, i.e., $x_1 > x_2$. Let us consider two different cases. (i) First case: $x_1 - x_2 < 1$. Suppose $f(x_1) \leq f(x_2)$. Then, by Lemma 3 and the definition of D_f,

there exist $h,e_1,e_2 \in L_c$ and $P''' \in \boldsymbol{P}$ such that $P'''(h|e_1) - P'''(h) = x_1$ and $P'''(h|e_2) - P'''(h) = x_2$. Clearly, if the latter equalities hold, then $P'''(h|e_1) > P'''(h|e_2)$. Thus, there exist $h,e_1,e_2 \in L_c$ and $P''' \in \boldsymbol{P}$ such that $C_{P'''}(h,e_1) = f(x_1) \leq f(x_2) = C_{P'''}(h,e_2)$ even if $P'''(h|e_1) > P'''(h|e_2)$, contradicting (C1). Conversely, (C1) implies that, for any $x_1,x_2 \in D_f$ such that $x_1 - x_2 < 1$, if $x_1 > x_2$ then $f(x_1) > f(x_2)$. (ii) Second case: $x_1 - x_2 \geq 1$. Given the definition of D_f, it is easy to show that $0 < x_1 - x_1/2 < 1$, $0 < x_1/2 - x_2/2 < 1$, and $0 < x_2/2 - x_2 < 1$ (here it is useful to note that, if $x_1 - x_2 \geq 1$, then $x_1 > 0 > x_2$). Relying on Lemma 3 as before and on the first case (i) above, we now have that (C1) implies $f(x_1) > f(x_1/2) > f(x_2/2) > f(x_2)$. Thus (C1) implies that, for any $x_1,x_2 \in D_f$ such that $x_1 - x_2 \geq 1$, if $x_1 > x_2$ then $f(x_1) > f(x_2)$. As cases (i) and (ii) are exhaustive, (C1) implies that, for any $x_1,x_2 \in D_f$, if $x_1 > x_2$ then $f(x_1) > f(x_2)$. By a similar argument, (C1) also implies that, for any $x_1,x_2 \in D_f$, if $x_1 = x_2$ then $f(x_1) = f(x_2)$. So, for (C1) to hold, it must be that, for any $h,e \in L_c$ and any $P \in \boldsymbol{P}$, $C_P(h,e) = f[P(h|e) - P(h)]$ and f is a strictly increasing function. □

CHAPTER 31

SELF-LOCATING CREDENCES

MICHAEL G. TITELBAUM

31.1 The Diachronic Problem

Just as an agent can be more or less confident that straw can be spun into gold, she can be more or less confident that today is Tuesday or that she is standing on Mount Hood. So besides representing an agent's degrees of belief about what the world is like, we might want to represent her degrees of belief about where (or when, or who) she is in the world.

Doing so may require us to reconsider the objects over which credences are distributed. Suppose, for example, we have been taking credences to be distributed over propositions defined as sets of possible worlds. To accommodate self-locating credences we might follow Lewis (1979) in distributing credences over propositions that are sets of *centered* worlds—where a centered world $\langle W, c \rangle$ is an ordered pair of a traditional possible world and a center.[1] Propositions composed of centered worlds can be sorted into two kinds: A proposition is uncentered if for any centered world $\langle W, c \rangle$ it includes, it also includes every other centered world indexed to the same W. Otherwise a proposition is centered. "Today is Tuesday" expresses a centered proposition, because it divides centered worlds indexed to the same traditional world from each other; "lead can be spun into gold" expresses an uncentered proposition. (Henceforth I will refer to traditional possible worlds simply as "worlds.")

Credences obeying the probability axioms can be distributed just as easily over propositions composed of centered worlds as they can over traditional propositions.[2] So this change in the objects of credence does not create any problems for Bayesians' usual synchronic constraints. Of course there is debate over whether propositions composed of centered

[1] A center is typically taken to be an ordered pair of a time and an individual. If we are allowing for the possibility of time-travel, a center may need to be a triple of a time, an individual, and a location. This allows for the possibility that having time-traveled, the same individual might be in two locations at the same time.

[2] The relevant technical point is that the set of propositions composed of centered worlds can form a sigma algebra just as easily as the set of propositions composed from standard worlds, which is what we require to assign a probabilistic distribution over the set.

worlds are the true objects of degrees of belief, but because that is the most popular position in the self-locating credence literature I will adopt it here. Most of the arguments in the objects-of-credence debate are familiar from the objects-of-belief debate, so credences don't add much that's new here.[3]

What is new is the havoc self-locating credences create for Bayesians' traditional diachronic constraint: updating by conditionalization. The current consensus in the self-locating credence literature is that obtaining a general updating scheme for degrees of belief in both centered and uncentered propositions requires us to alter (or at least supplement) conditionalization in some way.

To see why, suppose we have an agent who is currently certain that it is Tuesday. Intuitively, it is apparent there are some things that agent could learn as time goes on that would make it rational for her to decrease her certainty in that proposition. (In fact, if she is carefully watching a clock she is certain is accurate, there are some things the agent could learn that would make her credence in that proposition reach zero!) We want our updating rule to capture such transitions. But if conditionalization says that upon learning proposition E, the agent should change her credence in the proposition T that it is Tuesday to

$$P_j(\mathrm{T}) = P_i(\mathrm{T} \,|\, \mathrm{E}) = \frac{P_i(\mathrm{T} \,\&\, \mathrm{E})}{P_i(\mathrm{E})} \tag{31.1}$$

(where P_i and P_j are her prior and posterior credences, respectively), we have a problem: Since $P_i(\mathrm{T}) = 1$, the probability calculus tells us there is no E such that $P_i(\mathrm{E}) > 0$ and $P_i(\mathrm{T} \,\&\, \mathrm{E}) \neq P_i(\mathrm{E})$. So for any E we feed into the conditionalization rule, it either will tell us that $P_j(\mathrm{T}) = 1$ or will give us no guidance as to what that credence should be (because it will set that credence equal to a fraction with denominator zero).[4]

One might attribute this problem to the agent's (perhaps unreasonable) initial certainty in T, but similar problems arise for Jeffrey-conditionalization updating regimes that honor Regularity by forbidding certainty in empirical propositions.[5] So instead of retreating to Jeffrey conditionalization, a number of authors have granted that an agent might reasonably be certain of T and have proposed new formal updating schemes to replace conditionalization. These updating schemes replicate the effects of conditionalization when

[3] Though see Chalmers (2011) for the suggestion that credence considerations add new teeth to Frege's Puzzle arguments against referentialism about content. For other discussions of the objects of credence see Stalnaker (2008) and Pust (2012). In Titelbaum (2013b) I suggest that Bayesians can remain neutral among competing views about the objects of credence by modeling degrees of belief using numbers assigned to natural-language sentences.

[4] One might avoid this problem by making the following moves: first, read conditionalization as asserting the first equality in Equation (31.1) but not the second; next, define conditional credences in some way other than as ratios of unconditional credences (see Kenny Easwaran's chapter (9) in this volume); and finally, tell a positive story about the values of credences conditional on propositions with unconditional-credence zero that yields substantive and plausible results for cases such as the T/E case under discussion and the cases that appear later in this essay. Chalmers (2011, section 13) floats something like this proposal, but I have never seen anyone develop it into a full-blown updating system.

[5] For discussion of the problems self-location causes Jeffrey conditionalization see Kim (2009) and Titelbaum (2013b, chapter 12).

self-location is not involved, but also yield plausible results when agents update centered propositions such as T.[6]

31.2 A Master Narrative

The main difficulty in describing these new updating schemes is that there are so many of them (more than a dozen at the time of writing!). Instead of working through the details of each, I am going to sort them into three groups. For each group, I will (at a high level of abstraction) describe the basic updating approach of the schemes in that group, then describe difficulties common to every scheme in the group.

To understand the differences beween these three sorts of schemes, consider the following story:

> *Rick*: So far no rain has fallen on the 4th of July, but Rick can see ominous clouds approaching on the horizon. So he assigns a 0.7 credence to the proposition that it rains today. Before the clouds arrive and settle the matter he falls asleep, and then wakes up the next day (certain that he's slept for exactly one night and it's now July 5th). To what proposition should Rick now assign a 0.7 credence?

The role of an updating scheme is to coordinate credences assigned at different times. Conditionalization, for instance, requires an agent to line up her unconditional credences at a later time with particular conditional credences assigned earlier on. The story about Rick above also asks a coordination question: it asks us to pick a proposition whose July 5th credence should align with Rick's July 4th credence that it would rain that day. There are many propositions Rick might entertain on July 5th that would suit this role, among them the propositions expressed by these sentences:

- It rained yesterday.
- It rained on July 4th.
- It rained *that day* (where "that day" refers *de re* to July 4th).

Everyone will agree that Rick should assign a 0.7 credence to each of those three propositions on July 5th. But each of the schemes I will discuss focuses on one of the three, builds a formalism that directly calculates Rick's July 5th credence in that proposition, then sets Rick's credences in the other two via some process downstream. *Shifting schemes*, for example, provide a diachronic rule that links Rick's July 5th credence in the proposition expressed by "It rained yesterday" to his July 4th credence in the proposition expressed by "It rains today." Rick's July 5th credences in the other two propositions are then determined using synchronic rules and his July 5th credence that it rained yesterday.[7] *Stable base schemes*, on the other hand, determine Rick's July 5th credence that it rained on the 4th

[6] To simplify our discussion of these schemes I will henceforth set aside the Regularity issue, allow rational agents to assign certainty to contingent propositions, and not concern myself with how the schemes examined could be extended to Jeffrey-style updates.

[7] This is easily done using a probability-calculus theorem I call "Substitution," which says that if an agent is certain two propositions have the same truth-value, one proposition can be substituted for the

of July diachronically from his July 4th credences, then use that July 5th credence to synchronically determine the others. *Demonstrative schemes* work with the *de re* proposition before approaching the others.[8]

In the sections that follow I will present (simplified) representative examples of each sort of scheme. We will see that having its primary focus on a different one of these propositions leaves each sort of scheme with a different sort of blind spot.[9]

31.3 Shifting Schemes

Shifting schemes focus on the fact that in the Rick story, Rick's July 4th 0.7 credence in the proposition expressed by "It rains today" should become a July 5th 0.7 credence in the proposition expressed by "It rained yesterday." Intuitively, the latter proposition is generated from the former by *shifting* the indexicals backward one day (and changing tenses appropriately). July 4th credences assigned to propositions about the current day are aligned with July 5th credences assigned to propositions about the previous day; "today" becomes "yesterday," "tomorrow" becomes "today," etc.[10]

There are various ways of formally implementing this idea; shifting schemes have been offered by Kim (2009), Meacham (2010), Schulz (2010), Schwarz (2012), and Santorio (ms). Here we'll take Kim's scheme as representative.

Kim introduces an "at" operator that maps propositions to propositions. Kim tells us that for proposition ϕ and context τ, " 'ϕ at τ' means that ϕ is true at τ" (Kim 2009, p. 302). For example, if RT is the centered proposition expressed by "It rains today," [RT at Monday] is the uncentered proposition that it rains on Monday.[11]

other *salva veritate* into any expression describing the agent's credences. (For more on Substitution—and a proof—see Titelbaum (2013b, chapter 3).) This theorem entails that if an agent is certain two propositions have the same truth-value, he must assign them the same credence. Since Rick is certain on July 5th that all three bulleted propositions have the same truth-value, setting his credence in any one suffices to set his credence in the others.

[8] A great deal of the structure of this chapter—as well as many of the examples in it—were inspired by Branquinho (2008). Branquinho's main question is about the coordination of *thoughts*, not credences: in what circumstances can we say an agent has maintained the same thought (or belief) over time, or (approached slightly differently) what does it take to re-express the same thought one had earlier when that thought was earlier expressed in indexical terms? Branquinho's work is in turn a response to proposals in Evans (1981), Kaplan (1988b), and Kaplan (1988a). (I am grateful to Peter Ludlow for introducing me to Branquinho's work and for discussions of it.)

[9] There are a number of discussions of diachronic self-locating credence constraints that, while excellent, do not put forward a comprehensive formal scheme for updating self-locating degrees of belief. (See, for example, Bostrom (2007), Bradley (2010), Cozic (2011), Bradley (2011), Horgan and Mahtani (2013), and Manley (ms).) For reasons of space I will not discuss those here.

[10] From a completely unscientific study of exactly one subject, I can report that even relatively new language users are capable of shifting indexicals in this fashion. When her mother told my three-year-old, "*Your father* can give *you* some of *my* chocolate," she immediately turned to me and reported, "*You* can give *me* some of *mom's* chocolate."

[11] Kim doesn't define the "at" operator much more precisely than this, but we can give a simple definition that works at least for the case in which τ is a specific center. In this case the proposition [ϕ at

Now consider cases in which an agent learns nothing between times t_i and t_j except that the time has advanced from t_i to t_j. In these cases, Kim argues that the agent should update according to this rule:

$$P_j(X) = P_i(X \text{ at } t_j)$$

The idea here is that if the agent learns nothing between t_i and t_j about how events unfold in the world, his credence at t_j that X is true then should be the credence he had at t_i that X would be true at t_j.

Since Rick learns nothing during our story except that some time has passed, this rule applies to his case. So if RY is the centered proposition expressed by "It rained yesterday," we have

$$P_{\text{July } 5}(\text{RY}) = P_{\text{July } 4}(\text{RY at July } 5) \tag{31.2}$$

The proposition yielded on the right-hand side by the "at" operator is the (uncentered) proposition that it rains one day before July 5th. But that, of course, is just the proposition that it rains on July 4th, which is captured on July 4th by RT.[12] So we have

$$P_{\text{July } 5}(\text{RY}) = P_{\text{July } 4}(\text{RT}) \tag{31.3}$$

This says that Rick's July 5th credence that it rained yesterday should equal his July 4th credence that it rains today. Kim's "at" operator formally does the job we do informally by shifting indexicals around—transforming "today" to "yesterday," "tomorrow" to "today," and so forth as the days roll by.

What about cases in which agents learn more than just that time has passed? The rule we've just seen generalizes to Kim's

Shifted Conditionalization If an agent learns proposition Y (and nothing stronger) between t_i and t_j, and is certain at t_j that it is t_j, then for any proposition X

$$P_j(X) = P_i(X \text{ at } t_j \,|\, Y \text{ at } t_j)$$

For example, suppose Rick's story were changed so that upon awakening on July 5th he learned that it had stormed the previous evening. Then Shifted Conditionalization would give us

$$P_{\text{July } 5}(\text{RY}) = P_{\text{July } 4}(\text{RY at July } 5 \,|\, \text{SY at July } 5) = P_{\text{July } 4}(\text{RT} \,|\, \text{ST}) \tag{31.4}$$

where SY is the proposition expressed by "It stormed yesterday evening" and ST is the proposition expressed by "It storms this evening." Equation (31.4) tells us that if Rick wakes up on July 5th and learns that it stormed the previous evening, his credence that it rained on July 4th should be just what his credence would have been on July 4th that it would rain *that*

$\tau]$ is the set of world-center pairs $\langle W, c\rangle$ such that $\langle W, \tau\rangle$ is a member of ϕ. This means that whenever τ is a specific center, $[\phi \text{ at } \tau]$ will be an uncentered proposition. (It also means that applying the "at" operator with a specific center to an uncentered proposition yields the same uncentered proposition.) Since our examples all concern day-long contexts, a day will count as a specific center for our purposes.

[12] More precisely, Rick is certain on July 4th that the proposition RT has the same truth-value as the proposition [RY at July 5]. So Substitution (see note 7 above) allows us to move from Equation (31.2) to Equation (31.3).

day had he learned on July 4th that it was going to storm *that* evening.[13] Which is exactly right.

Yet Kim's Shifted Conditionalization has a shortcoming, revealed by stories like this:

> *Rip van Winkle*: On the 4th of July Rip van Winkle has credence 0.7 in "It rains today." He falls asleep and wakes up some time later, wildly uncertain what day it is. To what proposition should he now assign a credence of 0.7?

Intuitively, shifting schemes want to answer this question by taking Rip's July 4th credence in the proposition expressed by "It rains today" and replacing "today" with some other indexical reflecting the relationship of July 4th to the day when Rip awakens. The trouble is that Rip isn't certain how many days he slept, so he doesn't know which indexical ("yesterday"? "two days ago"? "a week ago"?) is right for the job. To shift properly, an agent needs to be certain how the context in which he's assigning his current credences relates to the context in which he assigned the credences he's updating.

This problem is reflected formally in Shifted Conditionalization's requirement that the agent be certain at t_j that it is t_j. One could (and Kim does) suggest that if the agent is uncertain at t_j whether it is t_j or some other time t_k, the agent's t_j credence in X should be a weighted average of the credence she would assign were she certain it was t_j and the credence she would assign were she certain it was t_k, with the weights determined by her unconditional t_j credence that it is t_j or t_k (respectively). The trouble with this is that the agent's unconditional t_j credences about which day it is have to be set antecedently to applying this weighting rule, via some other device altogether. For an agent who is uncertain what time it is (or where she is, or who she is), the shifting scheme itself will provide few constraints on her credences in the relevant self-locating propositions. And so the shifting scheme will not be able to tell the agent which new credences to coordinate with her old ones. Shifting schemes are very little help telling Rip which proposition gets a 0.7 credence when he awakens.

31.4 Stable Base Schemes

When Rip awakens he may be rationally required to distribute his credences over the days it might be in a particular way. Before he fell asleep Rip may have assigned credences to uncentered propositions expressed by "When Rip falls asleep on July 4th he awakens on July 5th," "When Rip falls asleep on July 4th he awakens on July 6th," etc. These credences should presumably remain intact when Rip wakes up, which will drive him to assign particular credences to the centered propositions expressed by "Today is July 5th," "Today is July 6th," etc.

Shifting schemes are not designed to take advantage of these credences Rip assigned on July 4th; stable base schemes, on the other hand, are built entirely around an agent's credences in uncentered propositions. Stable base schemes rely on the fact that only centered

[13] Again, the last step in Equation (31.4) follows by Substitution (note 7).

propositions make trouble for conditionalization.[14] Their core idea is to focus on the set of uncentered propositions an agent entertains, updating the members of that *stable base* by standard conditionalization on the uncentered propositions the agent learns. Credences in centered propositions are then set by coordinating them with the uncentered distribution at a given time. Stable base schemes are offered by (Halpern 2005), (Meacham 2008), (Titelbaum 2008), (Briggs 2010), and (Titelbaum 2013b).[15]

While Halpern's and Meacham's schemes are presented somewhat differently, they ultimately describe the same three-step process for generating a P_j distribution from one's earlier P_i assignments, a process Meacham calls "Compartmentalized Conditionalization":[16]

1. Temporarily consider your P_i distribution only over worlds.
2. Assign any worlds incompatible with your t_j evidence a credence of 0, then renormalize the remaining (non-zero) credences over worlds.
3. Now assign any centered worlds incompatible with your t_j evidence a credence of 0. Then take the credence assigned to each world and distribute it among the centered worlds indexed to that world compatible with your t_j evidence.[17]

Intuitively, Compartmentalized Conditionalization generates a new distribution over uncentered propositions by conditionalizing on the strongest uncentered proposition the agent learns (that's step (2)), then leaving that uncentered distribution intact while sorting out credences for particular centered worlds (step (3)). In Rip's case, Rip assigns a 0.7 July 4th credence to the uncentered proposition R4 that it rains on July 4th. He falls asleep, awakens, then applies Compartmentalized Conditionalization. Since Rip hasn't eliminated any worlds upon awakening, the conditionalization in step (2) does not alter his credences in any uncentered propositions. So he continues to assign R4 a credence of 0.7, exactly as he should.

[14] Titelbaum (2013b, chapter 8) argues that the true troublemakers for conditionalization are not centered propositions but instead "epistemically context-sensitive" propositions, where the epistemically context-sensitive/context-insensitive distinction cross-cuts the centered/uncentered distinction in some cases. We will ignore that complication here.

[15] Titelbaum (2013b) presents a version of the formal updating framework described in Titelbaum (2008) that has been altered in response to a counterexample from Moss (2012). (For an informal presentation of the new framework, see Titelbaum (2013a).) Briggs (2010) presents two formal updating schemes, one of which matches that of Meacham (2008). Meanwhile Meacham has abandoned the stable base scheme presented in Meacham (2008) in favor of the new shifting scheme of Meacham (2010), because of counterexamples he describes in this later work.

[16] It's easiest to explain Compartmentalized Conditionalization in terms of an agent's "credence distribution over worlds" and her "distribution over centered worlds." By "credence in a world W" I mean credence in the uncentered proposition containing all and only the centered worlds indexed to W; by "credence in a centered world $\langle W, c\rangle$" I mean credence in the proposition containing only that world. Once one has a credence distribution over centered worlds the probability axioms generate unique credences for all other propositions.

[17] Halpern and Meacham each discuss specific proposals for apportioning one's credence in a world among the centered worlds indexed to it—for example, one could apply an indifference principle (à la Elga (2004)) to distribute a world's credence equally among its centered worlds compatible with the agent's t_j evidence. But the details of those proposals are irrelevant to the points I will make here.

Compartmentalized Conditionalization can also take advantage of any credences Rip assigned on July 4th about how long he'd be asleep. If Rip was 0.5 confident on July 4th in the uncentered proposition that he'd sleep until July 5th, then step (2) has him retain this credence upon awakening and step (3) makes this a 0.5 credence that it's now July 5th.[18] We could also, if we like, add to Rip's story that he gains some uncentered information upon awakening—perhaps he learns that it stormed the evening of July 4th. If we call that uncentered proposition S4, step (2) of Compartmentalized Conditionalization would set Rip's new credence in R4 equal to $P_{\text{July 4}}(\text{R4} \mid \text{S4})$.

Under Compartmentalized Conditionalization all the changes to an agent's uncentered distribution happen in step (2), and such changes can be made only by ruling out worlds. So Compartmentalized Conditionalization is committed to the

Relevance-Limiting Thesis An agent should change her credence distribution over uncentered propositions only if her new evidence eliminates worlds.[19]

The Relevance-Limiting Thesis has some intuitive appeal—it suggests that if a learning episode leaves one's space of epistemically possible worlds intact, it should leave one's credence distribution over that space unaltered as well. But the following counterexample demonstrates that the Relevance-Limiting Thesis is false:

> *Mystery Bag*: You are one of ten people arranged in a circle in a room. A fair coin has been flipped to determine the contents of a bag: on heads the bag contains nine black balls and one white ball; on tails it's nine white and one black. The bag is passed around the room. Each person draws one ball, holds onto it, and passes the bag until it's empty. You can't see anyone else's ball, but the one you've drawn is black.

Everyone should agree that seeing your ball is black should increase your confidence that the coin came up heads; a standard Bayesian calculation shows your credence in heads increasing from 1/2 to 9/10 upon that discovery. But now let's add the wrinkle that in this story you have no qualitative way of discriminating between yourself and the other subjects in the room. We'll have to imagine that you don't know your own name (perhaps your memory has been erased, or perhaps you were raised by scientists and never given a name); you look exactly like everyone else in the room; the room is cylindrical so you can't describe yourself as, say, "The guy in the corner"; etc. Whatever science-fiction elements are needed to make this work, consider them added to the story.[20]

The point of this wrinkle is that when you see the black ball, your new evidence does not rule out any worlds you entertained earlier. You knew there would be at least one black ball in the bag no matter what the coin-flip outcome, so you knew all along there would exist one

[18] Notice, by the way, that even if Rip didn't have any July 4th credences about how long he'd be asleep, Compartmentalized Conditionalization could still find a proposition at the later time to coordinate with Rip's initial 0.7 credence in RT. That's because—unlike the shifting schemes—Compartmentalized Conditionalization doesn't rely on Rip's opinions about how long he's been asleep to locate that coordinating proposition.

[19] I introduced this thesis in a slightly different form in Titelbaum (2008). See Draper (2013) for further discussion.

[20] While I discovered the Mystery Bag example independently (and presented it in Titelbaum (2013b, chapter 10)), it is very similar to a counterexample to the Relevance-Limiting Thesis presented in Bradley (2011, section 9), which Bradley in turn attributes to Matt Kotzen.

person in the room who saw a black ball. You gain evidence upon seeing your ball, but it's the centered proposition expressed by "I see a black ball." And since such evidence enters into Compartmentalized Conditionalization only at step (3), it is incapable of changing your credence in any uncentered propositions. So Compartmentalized Conditionalization will leave your credence in the uncentered proposition that the coin came up heads unaltered by your new evidence, which is clearly the wrong result. This incorrect result will be duplicated by any updating scheme committed to the Relevance-Limiting Thesis.[21]

Most of the time, things we can pick out indexically we can also pick out qualitatively. Thus our credences in centered propositions are mirrored in uncentered propositions as well: on July 4th Rip assigns a credence of 0.7 not only to the centered proposition expressed by "It rains today" but also to the uncentered proposition expressed by "It rains on July 4th." But in some (perhaps highly artificial) cases we lack descriptive means of uniquely picking out the targets of indexicals. In those cases particular pieces of centered evidence gained may not be mirrored in uncentered propositions learned. In Mystery Bag, your inability to qualitatively pick yourself out means that the centered evidence that you've drawn a black ball isn't reflected in any uncentered proposition learned. Yet that centered evidence can (and should) influence your credences in uncentered propositions like heads. Because relevant centered evidence may go unreflected in uncentered propositions, we cannot follow Compartmentalized Conditionalization in letting only uncentered propositions learned affect one's uncentered distribution.

Titelbaum (2013b) proposes an updating scheme (called CLF, for "Certainty-Loss Framework") that addresses this flaw by limiting the application of Compartmentalized Conditionalization. CLF allows an agent to update by Compartmentalized Conditionalization only if for each time and each centered proposition she entertains, the agent has some uncentered proposition she is certain at that time has the same truth-value as the centered proposition. This ensures that Compartmenalized Conditionalization will be applied only when every piece of centered evidence learned is mirrored for the agent in some uncentered proposition; in such cases it seems safe to let the agent's uncentered distribution be driven entirely by the uncentered propositions learned.

CLF prevents Compartmentalized Conditionalization from yielding incorrect verdicts for cases like Mystery Bag. However, it introduces a new set of blind spots, typified by the following story:

> *Recurring Ron*: Ron has a serious sleep problem: He keeps falling asleep for a random, unpredicatble number of days. He's lost track of how many times this has happened, and each time he awakens he is uncertain for how many days he slept. On one particular awakening Ron remembers of an earlier day he was awake that he was 0.7 confident it would rain that day. But Ron isn't certain that day occurred during his most recent awakening, and he's pretty

[21] Notice that if the agent in Mystery Bag *does* have a way of picking himself out descriptively, he gains uncentered evidence (something like "Bob Jones picks a black ball") over the course of the story. In that case, Compartmentalized Conditionalization will correctly lead him to increase his confidence in heads. The fact that adding seemingly irrelevant distinguishing information can change Compartmentalized Conditionalization's prescriptions is what leads Meacham to criticize it in Meacham (2010). But my complaint here is not about the *difference* between Compartmentalized Conditionalization's recommendations in the two cases—it's about the fact that Compartmentalized Conditionalization gets the anonymous case *wrong*.

> confident that wasn't the only day he's ever been 0.7 confident of rain. To what proposition should Ron now assign a credence of 0.7?

Stable base schemes try to cash out an agent's credences entirely in uncentered terms, which are straightforwardly manipulable by conditionalization. But when an agent has no qualitative way of identifying a day (or a place, or a person)—as Ron lacks a unique way of qualitatively describing the earlier day on which he assigned a 0.7 confidence to rain—such schemes fall short. CLF falls silent on Recurring Ron, identifying no coordinating proposition to which Ron should assign a 0.7 credence on the new day.[22]

31.5 Demonstrative Schemes

Of course there is a way for Ron to pick out the day in question; as long as he remembers it, he can refer to it *de re* as "that day." So there is a coordinating proposition available to Ron upon awakening: he can be 0.7 confident that it rained *that day*.

Demonstrative reference capabilities provide the core of updating schemes proposed by Moss (2012) and Stalnaker (2008).[23] I'll focus on the latter here. Stalnaker's key move is to reject the distinction with which this chapter began, the distinction between beliefs about what world one is in and beliefs about where (or when, or who) one is in that world. As he puts it, "Belief about where one is in the world is always also belief about what world one is in." (p. 55) When I believe that it is raining today, I also believe *of* the current day that it is rainy. So my belief that it's raining today is accompanied by a belief that rules out particular worlds, worlds in which *this* day is not a rainy one. And notice that this will be true even if I lack a qualitative, non-demonstrative way of uniquely picking out the current day. As Stalnaker sees it, centered information is *always* mirrored in uncentered propositions.[24]

To make this work, we need to understand worlds—traditional worlds, the kinds to which centers are *added* to make centered worlds—in a way that incorporates demonstrative information. (So that we can distinguish between worlds in which it rains *this* day and worlds in which it doesn't.) Stalnaker has technical proposals for how to do this; assessing those proposals goes far beyond the scope of this chapter.[25] But notice that

[22] Titelbaum (2013b) proposes a workaround for this problem: We introduce a nearby story in which Ron can qualitatively pick out the relevant day, derive verdicts for that story, then argue that this story is enough like the original that our verdicts should apply there too. But this is still a workaround—it isn't a *direct* application of CLF to the Recurring Ron case.

[23] The updating scheme in Santorio (ms) incorporates *de re* elements, but is still essentially a shifting scheme.

[24] Stalnaker might be uncomfortable putting the point in "centered" vs. "uncentered" terms, in part because his own formalism employs centers in a different way from Lewis's. Moss prefers to discuss "*de se*" vs. "*de dicto*" propositions. At times in Stalnaker (2008, chapter 3), he puts the point I'm trying to make in terms of whether it's possible for an agent to learn something without gaining any "objective" information.

[25] It's also beyond my ambit to adjudicate between two rival proposals for understanding demonstrative reference: one that interprets demonstratives as *de re* expressions vs. another that interprets them as something like abbreviated indexical expressions of the form "the day I am thinking about right now." Since I am analyzing Stalnaker's updating scheme I will follow him in adopting the *de re* interpretation.

if Stalnaker's right—if centered propositions always have uncentered reflections—then Compartmentalized Conditionalization always applies. Moreover, all the action in generating updated credences happens in Compartmentalized Conditionalization's step (2). Uncentered propositions get their updated credence assignments at that stage, which can then immediately be copied to their centered correlates.[26] And since step (2) is just good, old-fashioned conditionalization, Stalnaker suggests that conditionalization is the only updating rule we need.

But even with demonstrative contents taken into account, conditionalization won't suffice to coordinate every credence. Consider the following case:

> *Roger Foretold*: On July 4th Roger knows that he's about to begin an extended period of sleepings and awakenings (of the kind experienced by Recurring Ron). This process begins, and some number of days later Roger finds himself awake, uncertain which awakening it is or how long he's been asleep. On this awakening Roger looks out the window, sees clouds, and becomes 0.7 confident it will rain. With which of Roger's July 4th credences is this 0.7 credence coordinated?

This story asks a slightly different question from our previous ones: here we have a later credence value and want to know with which earlier credence it's coordinated. But if our updating rule is just conditionalization, that's a valid question. On a conditionalization updating scheme every future credence you assign—even after learning evidence you previously lacked—is coordinated with some credence you assigned in the past. So there ought to be some conditional credence Roger assigns on July 4th—it will rain on such-and-such day conditional on its looking cloudy out my window on such-and-such day—that, by conditionalization, becomes his 0.7 credence in rain on the later awakening.

This causes a problem for Stalnaker because conditionalization always moves from a credence distribution over a particular space to another distribution *over the same space*. For Roger to take his July 4th credences, update them by conditionalization upon seeing clouds, and wind up with credences about whether it will rain today, there have to be propositions he assigned credences to on July 4th that he can on this awakening recognize as being about the current day. But there are no such propositions—there is no available way for Roger to fill in the "such-and-such"es in the previous paragraph. On July 4th Roger might consider whether it will rain tomorrow, but when he awakens he is unsure whether it currently is the day he referred to as "tomorrow" on July 4th. On July 4th Roger might also consider the prospects for rain on particular days picked out by name ("July 5th," "July 6th," etc.) or by qualitative description ("the day I first awaken"), but again Roger doesn't know which of those days it is when he awakens. The demonstrative scheme was supposed to solve these lack-of-description problems using *de re* references; Recurring Ron couldn't describe his earlier awakening qualitatively or relate it to the current time indexically, but he could always refer to it as "that day." But Roger can't do that on July 4th, because you can't refer *de re* to the future. Demonstrative reference requires causal contact, which notoriously works in only one temporal direction.

Demonstrative updating schemes rely on uncentered (or "*de dicto*", or "objective") propositions invoked by demonstrative reference. As we causally interact with new objects

[26] Since the agent is certain a centered proposition is true just in case its uncentered correlate is, Substitution copies his unconditional credence in the latter back to the former.

we gain new referential abilities, allowing us to assign credences to uncentered propositions we could not entertain before (or perhaps the same propositions under new modes of description). But conditionalization alone cannot tell us how to update credences from a previous, smaller epistemic space to a newly-expanded one.[27] So demonstrative schemes fall short in cases like Roger Foretold.[28]

31.6 The Sleeping Beauty Problem

I've now presented a number of rival updating schemes, organized into three groups. It's important to understand that members of one group need not see members of other groups as getting things *wrong*. A demonstrative schemer whose updates are driven by demonstrative reference will certainly not *deny* the shifter's result that on July 5th Rick should be 0.7 confident that it rained yesterday. Members of each group simply think that by approaching self-locating update in their particular fashion they can achieve further results—fill in further blind spots—unobtainable on another approach. In fact, the most heated debates about on-the-ground results are often between members of the same group. Those debates often focus on

> *The Sleeping Beauty Problem*: A student named Beauty arrives on Sunday to volunteer for an experiment. She will be put to sleep on Sunday night, then the experimenters will flip a fair coin. If it comes up heads, they will awaken her on Monday morning, chat with her for a bit, then put her back to sleep. If the coin comes up tails, they will engage in the same Monday process, then erase all her memories of her Monday awakening, awaken her on Tuesday morning, chat with her for a bit, then put her back to sleep.
>
> Beauty is told all of this information, then put to sleep. She awakens on Monday morning, but because of the possibility of memory erasure is uncertain whether it is Monday. At that point, how confident should she be that the coin came up heads?

Elga (2000), who adapted this problem from Piccione and Rubinstein (1997) and introduced it to the philosophical literature, argued that the answer is $1/3$. Lewis (2001) responded that the answer is $1/2$. Since most parties to the resulting controversy recognized that Beauty's only new evidence on Monday morning is self-locating (something like "It is now Monday or Tuesday"), this problem spurred much of the literature on self-locating update.[29]

[27] Weatherson (2011) makes a similar criticism of Stalnaker's scheme, prompting Stalnaker (2011) to concede that conditionalization by itself will not generate all the updates he wants.

[28] Moss's updating rule is not just traditional conditionalization, and is cleverly arranged so that demonstrative terms need to be available only from the time they are introduced and thereafter. But her scheme still has trouble with stories in which an agent forgets information and so loses track of *de re* ascriptions made in the past.

[29] While Sleeping Beauty may seem like an odd science-fiction problem, it has been linked to a very real issue in the philosophy of quantum mechanics. It has been argued that any updating scheme yielding a $1/3$ answer to Sleeping Beauty must also yield the implausible result that every quantum experiment we ever conduct will favor Everettian ("many-worlds") interpretations of the quantum-mechanical formalism over standard ("Copenhagen") interpretations, regardless of the particulars of the experimental result. For citations on this discussion, and further links between Sleeping Beauty and broader philosophical issues, see Titelbaum (2013c).

One might wonder why the controversy over Beauty's heads credence—and indeed, the whole search for a self-locating updating scheme—couldn't have been settled by the sorts of Bayesian arguments that have been used to support diachronic rules before. For example, diachronic Dutch Books and minimizing-expected-inaccuracy arguments have been invoked to support Conditionalization, Jeffrey Conditionalization, the Reflection Principle (van Fraassen 1995), and so on.[30] Interestingly, one can generate Dutch Book arguments and minimizing-inaccuracy arguments for each solution to the Sleeping Beauty Problem. Our understanding of how such arguments apply to cases involving self-location has not advanced far enough to settle the Sleeping Beauty dispute.[31]

Sleeping Beauty also lies within the blind spots of *all three* sorts of updating scheme. When Beauty awakens on Monday she is uncertain what day it is, making trouble for shifting schemes. She also does not have a unique qualitative description for the current day, undermining stable base approaches. Finally, Beauty does not on Monday have a description also available to her on Sunday night (not even a demonstrative one) that she is certain picks out the current day, so she cannot generate current-day credences by conditionalizing the possibility space that was available to her then.[32]

Each of the updating approaches we have considered seems reasonable as far as it goes, but gives out at some point. The creators of various schemes recognize those schemes' limitations and in many cases have proposed workarounds, some of which suggest answers to the Sleeping Beauty Problem. But few of those schemes generate a specific solution to the problem by a simple, direct application of their formalism. Since each sort of formalism represents a different intuitive approach to coordinating credences over time, their failures suggest that our standard, intuitive approaches to reasoning about self-location may be inadequate to settle Sleeping Beauty. Perhaps it is no coincidence that despite scores of proposed formal and informal solutions, the Sleeping Beauty Problem remains controversial.

[30] Because the Reflection Principle is tightly tied to conditionalization, centered propositions cause just as much trouble for the former as the latter. Arntzenius (2003) describes a number of counterexamples to Reflection that involve self-location, including the Sleeping Beauty Problem. Elga (2007) responds with a modification to Reflection that essentially implements a shifting scheme. See also Schervish, Seidenfeld, and Kadane (2004) on this topic.

[31] For the Dutch Book Sleeping Beauty discussion see Arntzenius (2002), Hitchcock (2004), Vineberg (ms), Bradley and Leitgeb (2006), Bostrom (2007), Draper and Pust (2008), Lewis (2010), Briggs (2010), Ross (2010), and Peterson (2011). For minimizing expected inaccuracy in Beauty see Kierland and Monton (2005), Briggs (2010), and Pettigrew (ms).

[32] It may seem suspicious that Sleeping Beauty (and some of the stories we considered earlier, such as Recurring Ron and Roger Foretold) involve memory loss, a phenomenon that is known to cause trouble for traditional Bayesian updating rules. (Arntzenius went so far as to assert at one point (Arntzenius 2002) that "self-locating learning plays no relevant role in the Sleeping Beauty case. The real issue is how one deals with known, unavoidable, cognitive malfunction." He backed off from that position in Arntzenius (2003).) Authors such as Meacham (2010), Moss (2012), and Titelbaum (2013b) have explicitly worked mechanisms for modeling memory loss into their updating schemes, but these maneuvers do not remove the blind spots I described for those approaches.

Acknowledgments

I am grateful to participants in the 2011 Rio Workshop on *de se* Attitudes for discussion of this chapter, to the editors of this volume for helpful comments, and to Olav Vassend for assistance keeping up with the literature.

References

Arntzenius, F. (2002) Reflections on Sleeping Beauty. *Analysis*. 62. pp. 53–62.

Arntzenius, F. (2003) Some problems for conditionalization and reflection. *The Journal of Philosophy*. 100. pp. 356–70.

Bostrom, N. (2007) Sleeping Beauty and self-location: A hybrid model. *Synthese*. 157. pp. 59–78.

Bradley, D. (2010) Conditionalization and belief *De Se*. *Dialectica*. 64. pp. 247–50.

Bradley, D. (2011) Self-location is no problem for conditionalization. *Synthese*. 182. pp. 393–411.

Bradley, D. and Leitgeb, H. (2006) When betting odds and credences come apart: More worries for Dutch Book arguments. *Analysis*. 66. pp. 119–27.

Branquinho, J. (2008) On the persistence of indexical belief. In *Proceedings of the XXII World Congress of Philosophy*. Vol. 39. pp. 21–30.

Briggs, R. (2010) Putting a value on Beauty. In Gendler, T. S. and Hawthorne, J. (eds.) *Oxford Studies in Epistemology*. Vol. 3. pp. 3–34. Oxford: Oxford University Press.

Chalmers, D. (2011) Frege's puzzle and the objects of credence. *Mind*. 120. pp. 587–635.

Cozic, M. (2011) Imaging and Sleeping Beauty: A case for double-halfers. *International Journal of Approximate Reasoning*. 52. pp. 137–43.

Draper, K. (2013) The evidential relevance of self-locating information. *Philosophical Studies*. 166. pp. 185–202.

Draper, K. and Pust, J. (2008) Diachronic Dutch Books and Sleeping Beauty. *Synthese*. 164. pp. 281–87.

Elga, A. (2000) Self-locating belief and the Sleeping Beauty problem. *Analysis*. 60. pp. 143–7.

Elga, A. (2004) Defeating Dr. Evil with self-locating belief. *Philosophy and Phenomenological Research*. 69. pp. 383–96.

Elga, A. (2007) Reflection and disagreement. *Noûs*. 41. pp. 478–502.

Evans, G. (1981) Understanding demonstratives. In Parret, H. and Bouveresse, J. (eds.) *Meaning and Understanding*. Berlin: W. de Gruyter.

Halpern, J. Y. (2005) Sleeping Beauty reconsidered: Conditioning and reflection in asynchronous systems. In Gendler, T. and Hawthorne, J. (eds.) *Oxford Studies in Epistemology*. Vol. 1. pp. 111–42. Oxford: Oxford University Press.

Hitchcock, C. R. (2004) Beauty and the bets. *Synthese*. 139. pp. 405–20.

Horgan, T. and Mahtani, A. (2013) Generalized conditionalization and the Sleeping Beauty problem. *Erkenntnis*. 78. pp. 333–51.

Kaplan, D. (1988a) Afterthoughts. In Almog, J., Perry, J., and Wettstein, H. (eds.) *Themes from Kaplan*. pp. 565–614. Oxford: Oxford University Press.

Kaplan, D. (1988b) Demonstratives: An essay on the semantics, logic, metaphysics, and epistemology of demonstratives and other indexicals. In Almog, J., Perry, J., and Wettstein, H. (eds.). *Themes from Kaplan*. pp. 481–563. Oxford: Oxford University Press.

Kierland, B. and Monton, B. (2005) Minimizing inaccuracy for self-locating beliefs. *Philosophy and Phenomenological Research*. 70. pp. 384–95.
Kim, N. (2009) Sleeping Beauty and shifted Jeffrey conditionalization. *Synthese*. 168. pp. 295–312.
Lewis, D. (1979) Atittudes *de dicto* and *de se*. *The Philosophical Review*. 88. pp. 513–43.
Lewis, D. (2001) Sleeping Beauty: Reply to Elga. *Analysis*. 61. pp. 171–6.
Lewis, P. J. (2010) Credence and self-location. *Synthese*. 175. pp. 369–82.
Manley, D. (ms). On being a random sample. Unpublished manuscript.
Meacham, C. J. G. (2008) Sleeping Beauty and the dynamics of de se beliefs. *Philosophical Studies*. 138. pp. 245–70.
Meacham, C. J. G. (2010) Unravelling the tangled web: Continuity, internalism, non-uniqueness and self-locating beliefs. *Oxford Studies in Epistemology*. 3. pp. 86–125.
Moss, S. (2012) Updating as communication. *Philosophy and Phenomenological Research*. 85. pp. 225–48.
Peterson, D. (2011) Qeauty and the books: A response to Lewis's quantum Sleeping Beauty problem. *Synthese*. 181. pp. 367–74.
Pettigrew, R. (ms). Self-locating beliefs and the goal of accuracy. Unpublished manuscript.
Piccione, M. and Rubinstein A. (1997) On the interpretation of decision problems with imperfect recall. *Games and Economic Behavior*. 20. pp. 3–24.
Pust, J. (2012) Conditionalization and essentially indexical credence. *Journal of Philosophy*. 109. pp. 295–315.
Ross, J. (2010) Sleeping Beauty, countable additivity, and rational dilemmas. *Philosophical Review*. 119. pp. 411–47.
Santorio, P. (ms). Cognitive relocation. Unpublished manuscript.
Schervish, M. J., Seidenfeld, T., and Kadane, J. (2004) Stopping to reflect. *Journal of Philosophy*. 101. pp. 315–22.
Schulz, M. (2010) The dynamics of indexical belief. *Erkenntnis*. 72. pp. 337–51.
Schwarz, W. (2012) Changing minds in a changing world. *Philosophical Studies*. 159. pp. 219–39.
Stalnaker, R. C. (2008) *Our Knowledge of the Internal World*. Oxford: Oxford University Press.
Stalnaker, R. C. (2011) Responses to Stoljar, Weatherson and Boghossian. *Philosophical Studies*. 155. pp. 467–79.
Titelbaum, M. G. (2008) The relevance of self-locating beliefs. *Philsophical Review*. 117. pp. 555–605.
Titelbaum, M. G. (2013a) *De Se* epistemology. In Feit, N. and Capone, A. (eds.) *Attitudes "De Se": Linguistics, Epistemology, Metaphysics*. pp. 307–25. Stanford, CA: CSLI Publications.
Titelbaum, M. G. (2013b) *Quitting Certainties: A Bayesian Framework Modeling Degrees of Belief*. Oxford: Oxford University Press.
Titelbaum, M. G. (2013c) Ten reasons to care about the Sleeping Beauty problem. *Philosophy Compass*. 8. pp. 1003–17.
van Fraassen, B. C. (1995) Belief and the problem of Ulysses and the Sirens. *Philosophical Studies*. 77. pp. 7–37.
Vineberg, S. (ms). Beauty's cautionary tale. Unpublished manuscript.
Weatherson, B. (2011) Stalnaker on Sleeping Beauty. *Philosophical Studies*. 155. pp. 445–56.

CHAPTER 32

PROBABILITY IN LOGIC

HANNES LEITGEB

This chapter is about probabilistic logics: systems of logic in which logical consequence is defined in probabilistic terms. We will classify such systems and state some key references, and we will present one class of probabilistic logics in more detail: those that derive from Ernest Adams' work.

32.1 Probability in Logic

Logic and probability have long been studied jointly: Boole (1854) is a classical example. If 'logic' is understood in sufficiently broad terms, then probability theory might even be subsumed under logic (as a discipline). In the words of Ramsey (1926, p. 157): 'the Theory of Probability is taken as a branch of logic'. John Maynard Keynes and Edwin Thompson Jaynes held similar views, and variants of the view were defended more recently, e.g., by Howson (2003) and Haenni (2005). In that sense, the probabilistic explication of the confirmation of hypotheses (as initiated by Carnap 1950) may, for example, be regarded as a kind of probabilistic (or inductive) logic; see Vincenzo Crupi and Katya Tentori's chapter on 'Confirmation Theory' (30) in this volume. On the other hand, if used in such a broad manner, the label 'probabilistic logic' is no longer particularly informative as far as its 'logic' component is concerned.

In this chapter, we will restrict the term 'logic' to logic *proper*: a *logic* or logical system is a triple of the form $\langle \mathcal{L}, \vDash, \vdash \rangle$, where (i) $\mathcal{L}$ is a formal language, (ii) $\vDash$ is a semantically (model-theoretically) specified relation of logical consequence defined for the members of $\mathcal{L}$, (ii) $\vdash$ is a proof-theoretically (in terms of axioms and rules) specified relation of deductive consequence for the members of $\mathcal{L}$. Ideally, $\vdash$ is sound with respect to $\vDash$ (that is, the extension of the $\vdash$ relation is a subset of the extension of the $\vDash$ relation), and $\vdash$ is complete with respect to $\vDash$ (the extension of the $\vDash$ relation is a subset of the extension of the $\vdash$ relation). However, not every logical system will satisfy both of these properties. Logic qua discipline is then the area in which logics in this sense are defined and in which they are studied systematically.

Now consider a logic in such a sense of the word: call the formal language $\mathcal{L}$ *for* which $\vDash$ and $\vdash$ are defined the 'object language', and call the language *in* which $\vDash$ and $\vdash$ are defined the

'metalanguage'. We can then define *probabilistic logics* to be precisely those logics $\langle \mathcal{L}, \models, \vdash \rangle$ for which the definition of $\models$ involves reference to, or quantification over, probability measures (which are then usually defined for the formulas in $\mathcal{L}$ or for subformulas thereof). And the area in which probabilistic logics in this sense are specified and described is probabilistic logic *as a discipline*. Probabilistic logic in this sense is the topic of this chapter. Reference to, or quantification over, probability measures on the metalevel will thus be a given in anything that follows.

For instance, in section 32.3 we will consider a definition of logical consequence for object language conditionals that will look like this:

- We say that

$$\{\varphi_1 \Rightarrow \psi_1, \ldots, \varphi_n \Rightarrow \psi_n\} \models \alpha \Rightarrow \beta$$

iff for all $\epsilon > 0$ there is a $\delta > 0$, such that for all probability measures P:
if for all $\varphi_i \Rightarrow \psi_i$ it holds that $P(\psi_i|\varphi_i) \geqslant 1 - \delta$, then $P(\beta|\alpha) \geqslant 1 - \epsilon$.

All of the technical details concerning this definition will be explained in section 32.3. What is relevant right now is just that this is a typical example of specifying a probabilistic logic: for the logical consequence relation $\models$ is determined semantically by quantifying over probability measures ('for all probability measures P').

As we are going to see, research in probabilistic logic even in this restrictive sense still cuts through various disciplines: theoretical computer science, artificial intelligence, cognitive psychology, and philosophy, especially, philosophical logic, philosophy of science, formal epistemology, and philosophy of language. However, the focus of this chapter will be on those aspects of probabilistic logic that seem most relevant philosophically. For instance, the logical consequence relation defined above extends nicely to a logico-philosophical theory of indicative conditionals in natural language, as developed by Adams (1975).

The rest of this chapter is organized as follows. In section 32.2 we will turn to a classification of probabilistic logics along two dimensions: (1) those which involve neither reference to, nor quantification over, probability measures on the object level, and (2) those which do involve such references on the object level. And within the second class, we will distinguish between (a) probabilistic logics which do involve *explicit* reference to, or quantification over, probability measures on the object level, and (b) those for which this kind of reference or quantification remains implicit. We will regiment some of the essential references on probabilistic logic into the resulting simple classification system. When doing so, we will refrain from going into formal details. And our selection of references will be, obviously, biased and incomplete. Section 32.3 will then be devoted to a concrete, and formally more detailed, case study of probabilistic logic(s): Ernest Adams' logic of high probability conditionals and some of its close relatives and variations. We have chosen this example because it may safely be called the most influential, and probably also the most innovative, instance of *probability in logic* in the philosophical corner of the subject matter. We will present six types of semantics for high-probability conditionals in that section. Although these semantics will look different initially, and although they are based on different motivations, they will be seen to determine one and the same deductive system of axioms and rules for conditionals: Adams' system P. Section 32.3 will be based partially on material from chapters 9-11 in Leitgeb (2004) (albeit with substantial revisions).

Some final remarks before we start classifying probabilistic logics: First of all, in probabilistic logic, probability measures are typically defined on formulas rather than on

sets (events) as standard probability theory would have it. We will state the essentials of such probability measures for formulas in section 32.3, but only for the very restrictive case of the language of propositional logic. For the extension to first-order languages, see Gaifman (1964), Scott and Kraus (1966), Fenstad (1967), Fagin (1976), Gaifman and Snir (1982), Nilsson (1986), Richardson and Domingos (2006) (for probabilities of first-order formulas as given by Markov networks), and (as far as inductive logic is concerned) Paris (2011), which may all be taken to develop probabilistic model theories for first-order formulas. Standard models or truth evaluations for formulas are thereby replaced by probability measures, and truth values for formulas are replaced by probabilities.

Secondly, throughout this chapter, the underlying base logic for probability measures will be assumed to be classical. E.g., one of the axioms for probability measures on formulas (see section 32.3) will demand the probability of all classical tautologies to be 1. But there are also probability measures for which classical logic is not presupposed in this way: see J. Robert G. Williams' chapter on 'Probability and Non-Classical Logic' (12) in this volume for more information.

Thirdly, probability measures can be interpreted in different ways: as an agent's rational degrees of belief in propositions, as objective non-epistemic chances of the occurrence of physical events, as statistical ideal long-term frequencies of properties applying to individuals in a certain domain, and more. Mostly, probabilistic logics are open to different such interpretations simultaneously, which is why we will not deal with the topic of interpreting probabilities very much, even though one of these interpretations is usually put forward as the 'intended' such interpretation (and in most cases that intended interpretation is the subjective 'Bayesian' one in terms of rational degrees of belief).

Fourthly, by turning our attention to logical consequence relations on formal object languages, we put to one side all theories that combine aspects of logic and probability in a different manner. In particular, there is a substantial literature on how to combine a logical account of (all-or-nothing) belief or acceptance with a probabilistic account of numerical degrees of belief, starting with Kyburg (1961), Hempel (1962), Levi (1967), Hilpinen (1968), and Swain (1970), through the more recent literature (for overviews see Foley 1993, Maher 1993, Christensen 2004, Huber and Schmidt-Petri 2009) to the most recent of such theories (e.g., Hawthorne and Makinson 2007, Sturgeon 2008, Wedgwood 2012, Lin and Kelly 2012, Leitgeb 2014, Buchak 2014, Ross and Schroeder 2014). Typically, these theories do *not* use formal languages (in the sense of formal logic) when stating the logical closure properties of belief, or the probabilistic axioms for degrees of belief, or principles of how belief relates to degrees of belief. Nor do they aim to define logical consequence relations for formal languages in probabilistic terms; which is why we will not cover these theories in this chapter.

32.2 The Classification of Probabilistic Logics

By our definition from the last section, a probabilistic logic $\langle \mathcal{L}, \models, \vdash \rangle$ includes a logical consequence relation $\models$ that is specified on the metalevel by referring to, or quantifying over, probability measures.

The first main decision point for probabilistic logics concerns the question of whether or not such a logic also involves reference to, or quantification over, probability measures on the *object level*:

1. Probabilistic logics which do *not* involve reference to, nor quantification over, probability measures on the object level:

   These are logical systems in which $\models$ is defined in probabilistic terms, but where the object language $\mathcal{L}$ itself (such as, e.g., the language of propositional logic) is not expressive enough to ascribe probabilities to formulas.

   One group of references in this category emerges from Popper (1955) who axiomatized primitive conditional probability measures for formulas *autonomously* from logic, that is, without presupposing (meta-)logical concepts such as *tautology* or *logical consequence* in the axioms for conditional probability themselves. Such conditional probability measures are not defined in terms of ratios of unconditional probabilities, as standard probability theory has it. That is why they can allow for a conditional probability $P(\beta|\alpha)$ to be defined even when $P(\alpha) = 0$ (see Halpern 2001, Makinson 2011, and Kenny Easwaran's chapter 'Conditional Probability' (9) in this volume for an overview). Although logical concepts are not used in their definition, these measures still end up being based on classical logic owing to the manner in which their axioms are set up. Indeed, in turn, it becomes possible now to *define logical concepts*, such as the relation of logical consequence, for the language of propositional logic *in purely probabilistic terms*. Popper's corresponding probabilistic account of logical consequence was extended later also to first-order languages by Field (1977), Leblanc (1979), van Fraassen (1981), and Roeper and Leblanc (1999), and to languages with modalities by Morgan (1982) and Cross (1993). For example, as far as the language $\mathcal{L}$ of propositional logic is concerned, Field suggests defining logical consequence probabilistically in the following manner: $\alpha_1, \ldots, \alpha_n \models \beta$ if and only if for every primitive conditional probability measure P on $\mathcal{L}$ that satisfies Popper's axioms, and for all formulas γ, it holds that $P(\beta|\gamma) \geqslant P(\alpha_1 \wedge \ldots \wedge \alpha_n|\gamma)$. Leblanc (1983) gives a simpler definition of consequence in terms of *un*conditional probabilities (which can be defined from conditional probabilities): $\alpha_1, \ldots, \alpha_n \models \beta$ if and only if for every probability measure P on $\mathcal{L}$, if for every α_i it holds that $P(\alpha_i) = 1$, then also $P(\beta) = 1$.

   The second group of references in this category has its source in Suppes (1966) who studied to what extent the probability of the conclusion of a logically valid argument may fall below the probabilities of the premises of the argument. It is easy to see that Suppes' observations can be turned into a probabilistic definition of $\models$ for the language $\mathcal{L}$ of propositional logic, as worked out in detail by Ernest Adams. Adams is also responsible for extending the account to conditionals $\alpha \Rightarrow \beta$ with a new primitive connective $\Rightarrow$ that is not definable by means of the connectives of propositional logic and which one may take to express high conditional probability. We will turn to Adams' work on high probability conditionals in more detail in section 32.3, but as far as the language $\mathcal{L}$ of propositional logic is concerned, logical implication for $\mathcal{L}$ may be defined probabilistically in the Suppes-Adams style as follows: $\alpha_1, \ldots, \alpha_n \models \beta$ if and only if for every probability measure P on $\mathcal{L}$ it holds that $P(\beta) \geqslant 1 - n + P(\alpha_1) + \ldots + P(\alpha_n)$. This consequence relation can then be shown to coincide extensionally with that of classical logic. Here is an example of how this result can be applied: since

$\alpha_1, \alpha_2 \vDash \alpha_1 \wedge \alpha_2$ in classical logic, it follows from applying the left-to-right direction of the equivalence above to the case of $n = 2$ that if $P(\alpha_1) \geqslant 1 - \epsilon$ and $P(\alpha_2) \geqslant 1 - \epsilon$, then $P(\alpha_1 \wedge \alpha_2) \geqslant 1 - 2\epsilon$ (and one can also show that this lower bound cannot be improved, unless additional information on the logical structure of α_1 and α_2 is available).

Now we turn to systems of probabilistic logic the object languages of which *are* expressive enough to ascribe probabilities to formulas.

2. Probabilistic logics which *do* involve reference to, or quantification over, probability measures on the object level:

The first class of such probabilistic logics concerns object languages that allow for ascribing probabilities to formulas *explicitly*:

A Probabilistic logics which do involve *explicit* reference to, or quantification over, probability measures on the object level:

By 'explicit' we mean: Either a sentential probabilistic operator or a probabilistic generalized quantifier is applied to a formula, or alternatively a probabilistic function symbol is applied to (the name of) a formula. The result of these applications is then combined somehow with expressions of the form '$= r$', '$\geqslant r$', or the like, where 'r' is a numeral denoting a real number in the unit interval. This leads to probabilistic formulas such as: '$P(\alpha) = r$' ('the probability of α is r') or '$P(\alpha) \geqslant r$' ('the probability of α is greater than or equal to r') or the like. A probability measure P can be said to satisfy such a formula, if interpreting the symbol 'P' by the measure P yields a true statement. Finally, logical consequence relations are defined for formal languages that include probabilistic formulas of such types. Usually, this is achieved by defining consequence in terms of truth preservation in all probability models for the object languages in question. Roughly: $\alpha_1, \ldots, \alpha_n \vDash \beta$ if and only if for every probability measure P on $\mathcal{L}$, if P satisfies each of $\alpha_1, \ldots, \alpha_n$, then P satisfies β. And deductive consequence relations may be defined which are then proven sound and, where possible, complete with respect to logical consequence. In a nutshell: this part of probabilistic logic deals with formalizations of the language of probability theory or various natural fragments thereof, such that the formalization pays off in terms of an improved control over the relations of logical and deductive consequence. This is in contrast to standard probability theory in which the entailment relation is usually left merely informal and implicit.

For instance, one way of presenting Hailperin's (1984, 1996, 2000) probabilistic logics is for them to involve object languages in which one can express that the probability of a formula is within a particular interval of real numbers. However, it is not yet possible to express in these languages the probabilities *of probability statements themselves*, that is, so-called second-order or higher-order probabilities, as for example '$P(P(\alpha) \in [r,s]) \in [r',s']$' ('the probability that the probability that α is in the interval $[r,s]$ is in the interval $[r',s']$'). Gaifman (1986) is the classical source for the logical treatment of higher-order probabilities, and most of the literature on probabilistic logics that allow for higher-order probabilities builds on it. (Such higher-order probability statements can be varied in lots of ways, so that

e.g. 'outer' and 'inner' occurrences of probabilistic symbols are assigned distinct interpretations; see e.g. Lewis 1980, van Fraassen 1995, Halpern 1991.)

Another landmark paper in that area is Fagin, Halpern, and megiddo (1990), who essentially logically formalize Nilsson's (1986) account of probabilistic reasoning on formulas (which had not been stated using a formal language). Their object language is even more expressive than what we dealt with before. For example, in their language one can say that the *weighted sum* of probabilities of finitely many formulas is greater than or equal to a certain real number, as in $a_1P(\alpha_1) + \ldots + a_nP(\alpha_n) \geqslant r$, where each of the α_i may again include logical connectives or P. The resulting language can also be extended to encompass Boolean combinations of such inequalities, inequalities for conditional probabilities, and first-order quantification of real numbers (based on the first-order theory of real closed fields). The authors provide semantic interpretations for these object languages. Relying on findings from linear programming, they determine sound and complete axiomatizations of the corresponding logical consequence relations, and they state NP-complete decision procedures for the correspondings satisfiability problems.

Here are some further closely related probabilistic logics in the same category: Frisch and Haddawy's (1988) object language is less expressive than Fagin, Halpern, and Megiddo's, although one can still say that the probability of a formula is in a certain interval of real numbers, and the corresponding probability operators can be nested again so that also higher-order probabilities can be ascribed. On the logical side, building on Gaifman's (1986) work, the semantics is set up so that Miller's principle—a typical instance of a higher-order so-called probabilistic reflection principle—is logically valid: $P(\alpha \mid P(\alpha) \in [r,s]) \in [r,s]$ ('the conditional probability of α, given that the probability of α is in the interval $[r,s]$, is in the interval $[r,s]$').

Heifetz and Mongin (2001) is another theory that employs a language less expressive than Fagin, Halpern, and Megiddo's but it comes with a special benefit: a less demanding fragment of arithmetic needs to be built into the corresponding axiomatic system. Speranski (2013) extends Fagin, Halpern, and Megiddo's account by adding also quantifiers over propositions.

The object language of Bacchus' (1990a, 1990b) probabilistic logic is even more expressive than Fagin, Halpern, and Megiddo's, at least as far as quantification is concerned, but there are also differences in terms of interpretation: while Fagin, Halpern, and Megiddo's probability measures are most easily interpreted as expressing subjective probabilities of closed sentences that determine sets of possible worlds, Bacchus also considers probability measures which are best interpreted as expressing statistical probabilities of open formulas that determine ensembles of individuals in a domain. (See also Hoover 1978 and Keisler 1985 for probabilistic logics with generalized quantifiers that concern probabilities of sets of tuples of individuals.) Such subjective and statistical probability measures can also be combined and expressed in one and the same logical system, as developed in Halpern (1990), Bacchus et al. (1996), and chapter 11 of Halpern (2003).

We should also mention some complexity results. We already mentioned Bacchus' (1990a, 1990b) system: his axiomatic system is complete with respect to models that are based on nonstandard probability measures. (More will be said about nonstandard probability in section 32.3.) But now consider systems for subjective

probability, or statistical probability, or both combined, where the object language includes: at least one probabilistic function symbol, the equality symbol, quantifiers (and at least one individual constant symbol). For any system of such type, Abadi and Halpern (1994) showed that if its set of logical truths is determined by models that involve only standard probability measures, then that set is not recursively axiomatizable anymore (unless further syntactic restrictions are invoked).

Fagin and Halpern (1994) extend the theory of Fagin, Halpern, and Megiddo (1990) in a different direction, by adding epistemic operators such as for knowledge, and Kooi (2003) and van Benthem, Gerbrandy, and Kooi (2009) further extend the account by invoking *dynamic* epistemic or probabilistic operators such as for knowledge change and probability change.

Finally, originating from a very different background—formal theories of truth and the study of semantic paradoxes (such as the famous Liar paradox)—Leitgeb (2012c) even allows for *type-free* probabilities: he presents different systems of probabilistic logic in which probabilities are ascribed to formulas that may speak about their *own* probabilities, such as a formula α that is provably equivalent to $P(\alpha) < 1$ (so that α may be said to express: *my probability is less than 1*). Type-free probability has been studied in most formal and philosophical detail by Campbell-Moore (2015b, 2016), who shows how one can turn existing formal theories of type-free truth into formal theories of type-free probability. Christiano et al. (unpublished) present an alternative theory of type-free probability, and Caie (2013) and Campbell-Moore (2015a) give reasons why one ought to be interested philosophically in type-free probability in that sense.

B Probabilistic logics which involve only *implicit* reference to, or quantification over, probability measures on the object level:

Whereas the previous category of probabilistic logics concerned ways of expressing probabilities on the usual numerical scale of concepts, the probabilistic logics in this category typically involve expressions for probabilities that merely occupy a categorical (all-or-nothing) or ordinal (comparative) scale of concepts. In the words of Halpern and Rabin (1987, p. 381): 'probability theory is not the only way of reasoning about likelihood'. The relations of logical consequence for the corresponding object languages are either defined in terms of truth preservation again—over probability measures themselves or over possible worlds models that are given a probabilistic interpretation—or not defined in terms of satisfaction at all. We have already encountered Adams' example of the latter kind in section 32.1, and we will return to this in more detail in section 32.3.

One group of such logical systems concerns formal languages with an 'it is (highly) probable that' operator. With it, one is able to express that $P(\alpha) \geqslant r$ (or maybe instead $P(\alpha) > r$) for a *fixed* real number threshold $\frac{1}{2} < r < 1$ that is not denoted explicitly in the object language. Hamblin (1959) was probably the first to study this (but still without iterations of the probability operator). Burgess (1969) presents a semantics for such an operator (in the strictly-greater-than version). He also presents sound, but not complete, axiomatizations of the corresponding logic even in the case in which nestings of the operator are allowed. So does Arló-Costa (2005), who suggests a neighborhood semantics for the operator. And Burgess (1969) gives decision procedures for the set of logically true (valid) formulas and the set of satisfiable formulas relative to a given threshold $\frac{1}{2} \leqslant r < 1$.

All of these logics are characterized by 'it is probable that α' and 'it is probable that β' *failing* to entail jointly 'it is probable that $\alpha \wedge \beta$', in line with the fact that the probability of a conjunction may fall below that of its conjuncts (as exemplified nicely in Kyburg's famous Lottery Paradox—see Wheeler 2007 for an overview). This is clearly in contrast with normal systems of modal logic, which are based on a possible worlds semantics rather than a neighborhood semantics. For according to them, 'it is necessary that α' and 'it is necessary that β' *do* jointly entail 'it is necessary that $\alpha \wedge \beta$'.

Halpern and Rabin (1987) develop yet another, though somewhat different, logical system for such an 'it is (highly) probable that operator'. And Terwijn (2005) studies a probabilistic logic the object language of which is that of first-order logic but where the truth condition of universally quantified formulas is given by a probabilistic threshold condition again.

The second group of references in this category concerns logics for an 'it is probabilistically certain that' operator by which one can express that $P(\alpha) = 1$. The system of Rescher (1962), in which the box or 'necessity' operator is interpreted in such a probabilistic manner, is an early example (but see also Hailperin 1937). The modal logic and semantics that emerge from this interpretation correspond to that of the standard modal system S5, in which nestings of the new operator are allowed and where 'it is probabilistically certain that α' and 'it is probabilistically certain that β' do jointly entail 'it is probabilistically certain that $\alpha \wedge \beta$'. Of course, this is just as it should be, as the axioms of probability do imply that the probability of a conjunction is 1 if the probability of each of its conjuncts is. Lando (2010) is a different, and more recent, example of a normal modal logic (in her case, S4) in which the box operator gets assigned a measure-theoretic interpretation (though a different one than Rescher's).

The next two groups of logical systems in the present class are extensions of the first and the second group, respectively, to probabilistic *conditional* operators $\Rightarrow$ in the object language, or to binary so-called *nonmonotonic consequence relations* $\mid\sim$ that are expressed metalinguistically but which may be viewed as corresponding to *sets* or *theories of probabilistic object-linguistic conditionals* closed under certain rules. In particular, Hawthorne (1996, 2007), Hawthorne and Makinson (2007), and Makinson (2012) study relations $\mid\sim$ between formulas in the language of propositional logic, such that $\alpha \mid\sim \beta$ if and only if $P(\beta|\alpha) \geqslant r$ (or $P(\alpha) = 0$). Here, P is again a given probability measure P, and r is a given real number threshold, so that $\frac{1}{2} < r < 1$, and the threshold is again not denoted explicitly in the object language. Arló-Costa and Parikh (2005) also determine nonmonotonic consequence relations probabilistically but they do so for the probability 1 case, such that $\alpha \mid\sim \beta$ if and only if $P(\beta|\alpha) = 1$. However, in their case P is assumed to be a primitive conditional probability measure as discussed briefly in the context of Popper's work in our first category of probabilistic logics from above. While

$$\frac{\alpha \mid\sim \beta, \alpha \mid\sim \gamma}{\alpha \mid\sim (\beta \wedge \gamma)}\text{(And)}$$

is logically valid in Arló-Costa and Parikh's system, it is invalid in Hawthorne and Makinson's system. If Arló-Costa and Parikh's logic for nonmonotonic consequence relations is reconstructed as a logic for conditionals—so that $\alpha \Rightarrow \beta$ expresses in the object language that $\alpha \mid\sim \beta$ (or $P(\beta|\alpha) = 1$) holds as expressed in the metalanguage—then the resulting *logical* consequence relation $\models$ for such conditionals is monotonic again, and it can be axiomatized in a sound and complete manner in terms of (Adams') logical system P in section 32.3 below. (For this to be the case it is crucial that Arló-Costa and Parikh assume 'P' to refer to a primitive conditional probability measure that satisfies Popper's axioms.) And if the logic of Hawthorne and Makinson is reconstructed as a logic of conditionals in a similar manner, then, metaphorically speaking, its axiomatization can be seen as the result of 'subtracting' the And rule above from the system P in section 32.3. However, it turns out to be quite difficult to state a sound and complete axiomatization for the logical consequence relation that is wanted. That relation $\models$ is given semantically by:

- $\{\varphi_1 \Rightarrow \psi_1, \ldots, \varphi_n \Rightarrow \psi_n\} \models \alpha \Rightarrow \beta$ iff for all P, for all $r \in [0,1]$: if for all $\varphi_i \Rightarrow \psi_i$ it holds that $P(\psi_i|\varphi_i) \geqslant r$ (or $P(\varphi_i) = 0$), then $P(\beta|\alpha) \geqslant r$ (or $P(\varphi) = 0$).

Hence, logical consequence corresponds to probability preservation above a threshold. Hawthorne and Makinson (2007) conjectured that their deductive system O was sound and complete (for Horn rules with finitely many premises). However, Paris and Simmonds (2009) proved it to be incomplete, while the infinite system of rules that Paris and Simmonds ultimately did prove to be sound and complete is highly complicated and not very intuitive.

Adams' logic of high probability conditionals, to which we will turn in more detail in the next section, lies somewhere in between Hawthorne and Makinson's and Arló-Costa and Parikh's accounts: the And rule that was mentioned above is logically valid in Adams' system (if stated for conditionals), while Adams' intended interpretation of $\alpha \Rightarrow \beta$ is that the conditional probability of β given α is *high, but not necessarily equal to 1*. In contrast to Hawthorne and Makinson's interpretation, the threshold of Adams' high-probability conditionals leaves the term 'high' only *vaguely* determined: no fixed real-number threshold is intended to be 'the' correct one (see the next section).

Finally, there are probabilistic logics which also belong to the present category but which represent probability measures in the object language differently from the logics discussed so far. For example, Segerberg (1971) and Gärdenfors (1975) study logical systems with an 'is at least as probable as' operator by which the laws of so-called qualitative probability (which originated with Bruno de Finetti) can be expressed in logical terms. Baltag and Smets (2008) develop logics for dynamic operators that represent probabilistic update in the object language. Yalcin (2010) presents a nice survey on probabilistic operators of various kinds from a philosophical and linguistic point of view. Yalcin's paper also includes further relevant references to logical studies of probabilistic operators on a categorical or ordinal scale.

32.3 A Case Study: Probabilistic Logics for High Probability Conditionals

In this final section we will study six distinct but extensionally equivalent semantics for 'high probability' conditionals, which all derive from Ernest Adams' work. Afterwards, we will turn to their axiomatic treatment.

32.3.1 Semantics for High Probability Conditionals

We need some preliminary definitions, before we can state the different versions of probability semantics for high probability conditionals.

First of all, for the rest of the chapter, let $\mathcal{L}$ be the formal language of standard propositional logic, except that $\mathcal{L}$ is restricted to only finitely many propositional variables $p_1, \ldots, p_n$. As far as logical symbols are concerned, we assume the standard connectives of classical propositional logic to be included in the vocabulary of $\mathcal{L}$: $\neg$, $\wedge$, $\vee$, $\rightarrow$ (for the material conditional), $\leftrightarrow$ (for material equivalence). So $\mathcal{L}$ contains formulas α such as $\neg p_1$, $p_2 \rightarrow (p_3 \vee p_4)$, $\neg(p_5 \wedge \neg p_5)$, and the like (assuming that $n \geqslant 5$), as usual.

Secondly, let W be the (finite) set of all classical truth value assignments to $p_1, \ldots, p_n$. More briefly, we shall speak of W as the set of all logically possible worlds over $\mathcal{L}$, since each single member of W determines uniquely a logically possible model or way of assigning truth values to the formulas in $\mathcal{L}$ in line with the usual semantic rules. If the model determined by w in W satisfies α, then we will denote this by: $w \vDash \alpha$. In the terminology of probability theory, W is going to function as the sample space of our probability measures; accordingly, the members of W may also be regarded as the possible outcomes of a random experiment.

Thirdly, with that set W being in place, call $\langle W, \wp(W), P\rangle$ a probability space (over W) if and only if (i) $\wp(W)$ is the power set over W (the set of all subsets of W), and (ii) P is a probability measure on $\wp(W)$, that is, $P : \wp(W) \rightarrow [0,1]$, $P(W) = 1$, $P(\varnothing) = 0$, and the axioms of finite additivity hold: for all $X, Y \subseteq W$, such that $X \cap Y = \varnothing$, it is the case that $P(X \cup Y) = P(X) + P(Y)$. (The axiom of so-called countable additivity or σ-additivity will not be assumed and will not play a role in any of the following.) Conditional probabilities can then be introduced by means of $P(Y|X) = \frac{P(X \cap Y)}{P(X)}$ in case $P(X) > 0$. In one of the semantic systems below we will actually deviate from this definition by allowing also for *non-standard real numbers* in the unit interval to be assigned by P; but in all of the other semantic systems we will stick to the definition just presented. The members of $\wp(W)$ will be called 'propositions'—W is the largest or 'tautological' proposition, $\varnothing$ is the least or 'contradictory' proposition—and thus probability measures in this sense assign probabilities to propositions and not (yet) to formulas. In standard probability theory, the members of $\wp(W)$ would rather be called 'events', but the difference is irrelevant really. (More importantly, standard probability theory allows for certain subsets of the sample space W not to be assigned probabilities at all; this will not be important either in what follows.)

Fourthly, although each P assigns probabilities to propositions, it may be used also, indirectly, to assign probabilities to formulas in $\mathcal{L}$ (and we will use the same function symbol

'P' for that purpose): for α in $\mathcal{L}$, let $[\alpha] = \{w \text{ in } W \mid w \vDash \alpha\}$. $[\alpha]$ is the set of worlds in which α is true, and we regard it as the proposition expressed by α. And for each α we can then define: $P(\alpha) = P([\alpha])$. Accordingly, for α, β in $\mathcal{L}$, define $P(\beta|\alpha) = \frac{P(\alpha \wedge \beta)}{P(\alpha)}$ in case $P(\alpha) > 0$. In order to simplify matters a bit later on, we will also regard $P(\beta|\alpha)$ to be well-defined, and indeed equal to 1, if $P(\alpha) = 0$. $P(\beta|\alpha)$ is the conditional probability that will be associated later with the high probability conditional $\alpha \Rightarrow \beta$.

Fifthly, following Ernest Adams' lead, we define the so-called uncertainty of β given α by means of $Unc(\beta|\alpha) = 1 - P(\beta|\alpha)$. $Unc(\beta|\alpha)$ will be the uncertainty associated with the high probability conditional $\alpha \Rightarrow \beta$.

Finally, let our conditional language $\mathcal{L}_{\Rightarrow}$ be the set of conditionals of the form $\alpha \Rightarrow \beta$ for which the antecedent α and the consequent β are formulas in $\mathcal{L}$. $\Rightarrow$ is a conditional connective that is not included in the vocabulary of $\mathcal{L}$, in particular, it is meant to differ from the symbol $\rightarrow$ for the material conditional. The intended interpretation of the new conditional $\alpha \Rightarrow \beta$ will be 'if α, then it is highly likely that β' or 'the conditional probability of β given α is high'. We want to leave open whether asserting such a conditional is meant to express the proposition *that the corresponding conditional probability is high* or whether asserting it merely expresses the corresponding *high conditional probability* in a more direct, non-propositional, 'expressivist' manner (for the difference between the two interpretations, see section 44 of Bennett 2003). Either way, we will call these conditionals 'high probability conditionals', so that 'their' probabilities are given in terms of their corresponding conditional probabilities, and it is these conditional probabilities that are taken to be high. As Lewis' (1976) showed in terms of his famous triviality theorems, and as subsequent work on the same topic made even clearer (such as that of Hájek 1989), this 'probabilities of conditionals are conditional probabilities' claim ought not to be understood in the way that the *unconditional* probability of the proposition expressed by $\alpha \Rightarrow \beta$ would be required to equal the *conditional* probability of β given α. This is because this would entail the underlying probability measure's being trivial as far as its range of possible numerical values is concerned, given only some very mild background assumptions. Instead, if one wants to speak of probabilities of conditionals at all, one should think of their probabilities as being *defined* as conditional probabilities without any assumption to the effect that probabilities of conditionals would also have to satisfy the axioms of *un*conditional probability. Also note that, syntactically, the members of our conditional language $\mathcal{L}_{\Rightarrow}$ are 'flat' in allowing neither for nestings of conditionals nor for the application of any of the connectives of classical propositional logic to conditionals. For instance, $\mathcal{L}_{\Rightarrow}$ does not include negations of conditionals.

When we are going to study logical consequence relations for this conditional language $\mathcal{L}_{\Rightarrow}$, we will focus on finite sets $KB_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$ of such conditionals, which will then function as finite sets of conditional premises or as finite (probabilistic) conditional knowledge bases (as theoretical computer scientists would say). We use the notation '$KB_{\Rightarrow}$', with the subindex '$\Rightarrow$', in order to signal that any such $KB_{\Rightarrow}$ is a set of conditionals. Although we do not include any 'factual', that is, non-conditional, formulas in $\mathcal{L}_{\Rightarrow}$ nor in any $KB_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$, for many applications one may think of conditionals $\top \Rightarrow \alpha$ with the tautological antecedent $\top$ as being logically equivalent to the factual formula α in $\mathcal{L}$. In particular, this makes good sense if one thinks of $\Rightarrow$ as representing the indicative 'if-then' in natural language, and, accordingly, Adams does treat α and $\top \Rightarrow \alpha$ as logically equivalent.

We are now ready to present six probabilistic semantics for high probability conditionals. Each semantics—except for the infinitesimal semantics—is based essentially on some probability semantics that had been suggested by Adams (see, e.g., Adams 1966, 1974, 1975, 1986, 1996, 1998, and Adams and Levine 1975). Adams' semantic systems were further refined and extended by Pearl (1988), McGee (1989), Lehmann and Magidor (1992), Edgington (1995), Goldszmidt and Pearl (1996), Schurz (1997, 1998), Snow (1999), Bamber (2000), Biazzo et al. (2002), Halpern (2003), Arló-Costa and Parikh (2005), and Leitgeb (2012a, b).

Each of the semantic systems below includes the definition of a logical entailment relation that holds between finite sets of high probability conditionals and further such conditionals. Each of these definitions will seem to be more or less plausible in itself, but they will all be based on different philosophical ideas and motivations: While semantics 2, 4, and 6 are defined in terms of *truth* preservation, semantics 1, 3, and 5 do not involve the notion of truth of a high probability conditional in a model at all. Whereas semantics 1 and 2 understand logical consequence *dynamically* in terms of 'the more likely the premises get, the more likely the conclusion gets', all the other semantics are static. Where semantics 3 and 5 concern the reliability of reasoning with conditionals, as they demand the probability of a conclusion not to drop too much below the probabilities of the premises, semantics 4 and 6 take a more idealized viewpoint by considering probabilistic orderings of worlds or infinitesimal probabilities. But, surprisingly, all of these definitions can be shown ultimately to determine (extensionally) *one and the same relation* of logical consequence for high probability conditionals, as we are going to see later. The resulting sound and complete deductive system of logical axioms and rules is Adams' logic P of conditionals, which therefore turns out to be robustly justified on quite diverse semantic grounds.

According to the first semantic system that we introduce, a set of high probability conditionals entails another high probability conditional if and only if: the higher the probabilities of the conditionals contained in the premise set, the higher also the probability of the conditional conclusion. This leads to a kind of 'continuity' semantics for high probability conditionals which, accordingly, employs an ϵ-δ-criterion:

Definition 1 *(Continuity Semantics for High Probability Conditionals)*

- *We say that*

$$KB_{\Rightarrow} \models_{cont} \alpha \Rightarrow \beta$$

iff for all $\epsilon > 0$ *there is a* $\delta > 0$*, such that for all probability measures P: if for all* $\varphi \Rightarrow \psi$ *in* $KB_{\Rightarrow}$ *it holds that* $P(\psi|\varphi) \geqslant 1 - \delta$*, then* $P(\beta|\alpha) \geqslant 1 - \epsilon$ *(that is: if* $P(\psi|\varphi)$ *is 'high' for all* $\varphi \Rightarrow \psi$ *in* $KB_{\Rightarrow}$*, also* $P(\beta|\alpha)$ *is 'high').*

It is well known that the definition of continuous functions over the reals can be stated either in terms of an ϵ-δ-criterion or in terms of the preservation of limits along sequences of real numbers. Similarly, the continuity semantics above also allows for a restatement in terms of a sequence semantics, where a sequence of probability measures is defined to satisfy a high probability conditional if the conditional probability associated with the conditional is identical to 1 'in the limit' of the sequence. Adams (1986, p. 277) hints at such a type of semantics in a footnote. Variants of such a sequence semantics—but defined on more

expressive languages than our simple $\mathcal{L}$—are employed by Halpern (2003) in his system of inductive reasoning for statistical and subjective probabilities, and by Leitgeb (2012a, b) in his probability logic for counterfactuals:

Definition 2 *(Sequence Semantics for High Probability Conditionals)*

- *A probabilistic sequence model* $\mathfrak{M}_{seq}$ *for high probability conditionals is a sequence* $(P_n)_{n\in\mathbb{N}}$ *of probability measures.*
- *Relative to a probabilistic sequence model* $\mathfrak{M}_{seq} = (P_n)_{n\in\mathbb{N}}$ *we can define:*

$$\mathfrak{M}_{seq} \models_{seq} \alpha \Rightarrow \beta$$

iff the real sequence $(P_n(\beta|\alpha))_{n\in\mathbb{N}}$ *converges, and*

$$\lim_{n\to\infty} P_n(\beta|\alpha) = 1$$

(that is: $P_n(\beta|\alpha)$ *'tends' towards* 1 *for increasing n).*
- $\mathfrak{M}_{seq} \models_{seq} KB_{\Rightarrow}$ *iff for every* $\alpha \Rightarrow \beta$ *in* $KB_{\Rightarrow}$ *it holds that* $\mathfrak{M}_{seq} \models_{seq} \alpha \Rightarrow \beta$.
- *We say that*

$$KB_{\Rightarrow} \models_{seq} \alpha \Rightarrow \beta$$

($KB_{\Rightarrow}$ *sequence-entails* $\alpha \Rightarrow \beta$*) iff*
for every probabilistic sequence model $\mathfrak{M}_{seq}$*:*
if $\mathfrak{M}_{seq} \models_{seq} KB_{\Rightarrow}$*, then* $\mathfrak{M}_{seq} \models_{seq} \alpha \Rightarrow \beta$.

Next, we turn to a semantics for high probability conditionals that does not involve anything like probabilities getting 'arbitrarily close to 1'. A set of high probability conditionals will instead be said to entail a high probability conditional if the uncertainty associated with the latter is smaller than or equal to the sum of the uncertainties of the conditionals contained in the premise set; that is, if the uncertainty of the conditional to be entailed is bounded additively by the uncertainties that are associated with the premise conditionals. In contrast with the two semantic systems from before, if a set of high probability conditionals entails another such conditional in this sense, there is always a lower bound for the probability that is associated with the conclusion, such that this lower bound can additionally be computed easily. As Schurz (1997) points out, the resulting entailment relation approximates the so-called 'quasi-tightness' property of inferences that was defined in Frisch and Haddawy (1994). The resulting uncertainty semantics, which had been introduced by Adams again, was taken up and defended for example by Edgington in her theories of indicative conditionals (Edgington 1995) and vague terms (Edgington 1996); similarly, Field (2009) models his account of how logical implication interacts normatively with degrees of belief after this kind of (Suppes-)Adams-style uncertainty semantics:

Definition 3 *(Uncertainty Semantics for High Probability Conditionals)*

- *We say that*

$$KB_{\Rightarrow} \models_{unc} \alpha \Rightarrow \beta$$

($KB_{\Rightarrow}$ uncertainty-entails $\alpha \Rightarrow \beta$) iff
for every probability measure P (and where a sum over an empty set of indices is defined to be 0):

$$P(\beta\,|\alpha\,) \geqslant 1 - \sum_{\varphi \Rightarrow \psi \in KB_{\Rightarrow}} Unc(\psi\,|\varphi\,),$$

that is,

$$Unc(\beta\,|\alpha\,) \leqslant \sum_{\varphi \Rightarrow \psi \in KB_{\Rightarrow}} Unc(\psi\,|\varphi\,)$$

(in words: $P(\beta\,|\alpha\,)$ is 'high' if the uncertainties $Unc(\psi\,|\varphi\,)$ are very 'low' for all $\varphi \Rightarrow \psi \in KB_{\Rightarrow}$; or: for all probability measures, it holds that the uncertainty associated with $\alpha \Rightarrow \beta$ is bounded from above by the sum of the uncertainties associated with the premises).

According to the next semantics, a high probability conditional is satisfied by a certain kind of probability measure that ranks worlds by polynomial 'orders of magnitude'. A high probability conditional is satisfied by such a probability measure if its associated conditional probability is of the *maximal* order of magnitude (compare Snow's 1999 'atomic bound probabilities' and Benferhat et al. 1997 on their so-called 'big-stepped probabilities'). The order-of-magnitude mapping may also be seen as a selection function in the sense of Stalnaker (1968) or as determining a special kind of sphere system of worlds in the sense of Lewis (1973). This explains the formal correspondence between the logic of high probability conditionals in the next subsection and Stalnaker's and Lewis' logical systems for counterfactuals. But the intended interpretation of Stalnaker's and Lewis' orderings in terms of similarity or closeness to the actual world differs from the purely probabilistic ordering of worlds below. Probabilistic order-of-magnitude models are also close to ranked models along the lines of Kraus, Lehmann, and Magidor (1990), and Lehmann and Magidor (1992)—which explains the formal correspondence between the logic in the next subsection with systems well-known from nonmonotonic reasoning. And these models are similar to ranking functions (or ordinal conditional functions) in the sense of Spohn (1988, 2012). This is what this order of magnitude semantics looks like in more formal terms:

Definition 4 *(Order of Magnitude Semantics for High Probability Conditionals)*

- *A probabilistic order-of-magnitude model $\mathfrak{M}_{om}$ for high probability conditionals is a bijective mapping $om : W \to \{0, \ldots, n-1\}$.*
  (So om is both one-to-one and onto: $om(w)$ is the 'probabilistic order of magnitude' of w. The cardinality of W, $card(W)$, is n.)
- *Relative to a probabilistic order-of-magnitude model $\mathfrak{M}_{om}(= om)$, and relative to some 'small' real number $v \in [0, 1]$ (say, $v < \frac{1}{2}$), we can define:*

  - *Let P_{om} be the unique probability measure that satisfies:*
    $P_{om}(\{w\}) = v^{om(w)}(1-v)$ *for* $om(w) < card(W) - 2$,
    $P_{om}(\{w\}) = v^{card(W)-1}$ *for* $om(w) = card(W) - 2$,

$P_{om}(\{w\}) = 0$ *for* $om(w) = card(W) - 1$.

- $$\mathfrak{M}_{om} \models_{om} \alpha \Rightarrow \beta$$

  iff $P_{om}(\beta\,|\alpha\,) \geqslant 1 - v$
  (that is: $P_{om}(\beta\,|\alpha\,)$ *is 'high' or corresponds to the highest order* $v^0(1 - v) = 1 - v$ *of magnitude).*

Note that whether $\mathfrak{M}_{om} \models_{om} \alpha \Rightarrow \beta$ *or not is actually independent of the exact choice of* v.

- $\mathfrak{M}_{om} \models_{om} KB_{\Rightarrow}$ *iff for every* $\alpha \Rightarrow \beta \in KB_{\Rightarrow}$ *it holds that* $\mathfrak{M}_{om} \models_{om} \alpha \Rightarrow \beta$.
- *We say that*

$$KB_{\Rightarrow} \models_{om} \alpha \Rightarrow \beta$$

($KB_{\Rightarrow}$ *order–of-magnitude-entails* $\alpha \Rightarrow \beta$*) iff*
for every probabilistic order-of-magnitude model $\mathfrak{M}_{om}$*:*
if $\mathfrak{M}_{om} \models_{om} KB_{\Rightarrow}$*, then* $\mathfrak{M}_{om} \models_{om} \alpha \Rightarrow \beta$.

The next semantics defines a set of high probability conditionals to entail a high probability conditional, if, whenever the conditional probabilities that are associated with the premises are 'close' to 1 (where the referent of 'close' is determined relative to the number of premises), the conditional probability that is associated with the conclusion, say, $\alpha \Rightarrow \beta$, is greater than $\frac{1}{2}$ and hence greater than the conditional probability that is associated with $\alpha \Rightarrow \neg\beta$. Since in any such case the set of β-worlds constitutes the 'majority' within the set of α-worlds (as measured by the probability measure in question), we call this the 'majority semantics'. Logical consequence given by this semantics therefore consists in the premises making the conclusion more likely than not:

Definition 5 *(Majority Semantics for High Probability Conditionals)*

- *Let* $KB_{\Rightarrow} = \{\varphi_1 \Rightarrow \psi_1, \ldots, \varphi_n \Rightarrow \psi_n\}$*:*
  We say that

$$KB_{\Rightarrow} \models_{maj} \alpha \Rightarrow \beta$$

($KB_{\Rightarrow}$ *majority-entails* $\alpha \Rightarrow \beta$*) iff*
for all probability measures P:
if $P(\psi_1|\varphi_1) > 1 - \frac{1}{2n}, \ldots, P(\psi_n|\varphi_n) > 1 - \frac{1}{2n}$*, then* $P(\beta|\alpha) > \frac{1}{2}$
(that is, if the premise probabilities are 'high', then $P(\beta|\alpha)$ *is greater than* $P(\neg\beta|\alpha)$*).*

The final semantics for high probability conditionals that we will discuss was suggested by Lehmann and Magidor (1992), pp. 48–53, and it is special insofar as it presupposes the *nonstandard* analysis of real numbers. Nonstandard analysis adds infinitely small numbers (the so-called 'infinitesimals') and infinitely large numbers to the standard set of real numbers. Apart from the introduction to nonstandard analysis that is contained in Lehmann and Magidor (1992) itself, brief but useful accounts of nonstandard analysis can

also be found in section 4.4 of Chang and Keisler (1990), and in more informal terms, in Adams (1998), pp. 253–6.

Definition 6 *(Infinitesimal Semantics for High Probability Conditionals)*

- *An infinitesimal probabilistic model $\mathfrak{M}_{inf}$ for high probability conditionals is a nonstandard probability measure $P : \wp(W) \to [0,1]^*$, that is, probabilities are nonstandard reals non-strictly between 0 and 1, such that $P(W) = 1$, $P(\emptyset) = 0$, and finite additivity is satisfied.*
- *Relative to an infinitesimal probabilistic model $\mathfrak{M}_{inf}(= P)$ we can define:*

$$\mathfrak{M}_{inf} \models_{inf} \alpha \Rightarrow \beta$$

iff $1 - P(\beta\,|\alpha)$ is infinitesimal, that is,
for all standard reals $\epsilon \in \mathbb{R}$ with $\epsilon > 0$: $1 - P(\beta\,|\alpha) < \epsilon$
(that is: $P(\beta\,|\alpha)$ is either identical to 1 or 'infinitely close' to 1).
- *$\mathfrak{M}_{inf} \models_{inf} KB_{\Rightarrow}$ iff for every $\alpha \Rightarrow \beta \in KB_{\Rightarrow}$ it holds that $\mathfrak{M}_{inf} \models_{inf} \alpha \Rightarrow \beta$.*
- *We say that*

$$KB_{\Rightarrow} \models_{inf} \alpha \Rightarrow \beta$$

($KB_{\Rightarrow}$ infinitesimally entails $\alpha \Rightarrow \beta$) iff
for every infinitesimal probabilistic model $\mathfrak{M}_{inf}$:
if $\mathfrak{M}_{inf} \models_{inf} KB_{\Rightarrow}$, then $\mathfrak{M}_{inf} \models_{inf} \alpha \Rightarrow \beta$.

This concludes our series of semantical systems for high probability conditionals.

We are now ready to turn to a comparison between these different versions of a high probability semantics in terms of their respective logical consequences relations. Surprisingly, the semantic systems that we presented in this subsection turn out to be *extensionally mutually equivalent* in the following sense (for the theorem, and for information on which proofs in the relevant literature the theorem is based, see Leitgeb 2004, pp. 177f):

Theorem 7 *(Equivalence of the Different Versions of Probability Semantics with respect to Entailment)*

Let $KB_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$, $\alpha \Rightarrow \beta \in \mathcal{L}_{\Rightarrow}$; the following claims are equivalent:

1. $KB_{\Rightarrow} \models_{cont} \alpha \Rightarrow \beta$
2. $KB_{\Rightarrow} \models_{seq} \alpha \Rightarrow \beta$
3. $KB_{\Rightarrow} \models_{unc} \alpha \Rightarrow \beta$
4. $KB_{\Rightarrow} \models_{om} \alpha \Rightarrow \beta$
5. $KB_{\Rightarrow} \models_{maj} \alpha \Rightarrow \beta$
6. $KB_{\Rightarrow} \models_{inf} \alpha \Rightarrow \beta$.

In the next subsection, we will determine the very consequence relation that corresponds to these semantic systems in proof-theoretic terms.

32.3.2 Proof Theory for High Probability Conditionals

Consider the following rules of inference for conditionals in $\mathcal{L}_{\Rightarrow}$ (where '$\vdash$' denotes the derivability relation of classical propositional logic):

- $\dfrac{}{\alpha \Rightarrow \alpha}$ (Reflexivity)
- $\dfrac{\alpha \vdash \beta, \beta \vdash \alpha, \alpha \Rightarrow \gamma}{\beta \Rightarrow \gamma}$ (Left Equivalence)
- $\dfrac{\gamma \Rightarrow \alpha, \alpha \vdash \beta}{\gamma \Rightarrow \beta}$ (Right Weakening)
- $\dfrac{(\alpha \wedge \beta) \Rightarrow \gamma, \alpha \Rightarrow \beta}{\alpha \Rightarrow \gamma}$ (Cautious Cut)
- $\dfrac{\alpha \Rightarrow \beta, \alpha \Rightarrow \gamma}{(\alpha \wedge \beta) \Rightarrow \gamma}$ (Cautious Monotonicity)

Note that Reflexivity is premise-free (so it is really an axiom scheme).

Kraus, Lehmann, and Magidor (1990, section 3), refer to the system of rules above as the system C of cumulative reasoning (although they spell things out in terms of nonmonotonic consequence relations rather than of conditionals). Cumulativity, that is, Cautious Cut and Cautious Monotonicity taken together, has been suggested by Gabbay (1984) to be a valid closure property of plausible reasoning: Cautious Monotonicity expresses that importing consequents (such as β) into an antecedent (so that α is turned into $\alpha \wedge \beta$) does not *subtract* from the original antecedent's (α's) inferential power. In turn, Cautious Cut expresses that importing consequents in this way does not *add* to the antecedent's inferential power either: for consider the denial of the conclusion. Then at least one of the two premises does not hold. If $\alpha \Rightarrow \beta$ does hold, so that β is a consequence of α, then we cannot infer γ by importation of that consequence.

Furthermore, we also consider the following rule:

- $\dfrac{\alpha \Rightarrow \gamma, \beta \Rightarrow \gamma}{\alpha \vee \beta \Rightarrow \gamma}$ (Disjunction)

The system that results from adding the Disjunction rule to system C is called *the system P of preferential reasoning* by Kraus, Lehmann, and Magidor (1990, section 5). This stronger system P is one of *the* standard systems of nonmonotonic logic, and it turns out to be sound and complete with respect to many different semantics of nonmonotonic logic (some of them are collected in Gabbay et al. 1994; see also Gärdenfors and Makinson 1994, Chapter 4.3 of Fuhrmann 1997, Benferhat et al. 1997, and Benferhat et al. 2000). Psychological findings, though still on a very preliminary level, indicate that P incorporates some of the rationality postulates governing human commonsense reasoning with conditionals (see Pfeifer and Kleiter 2005, 2010). P also coincides with the 'flat' fragment of Stalnaker's and Lewis' logic(s) for counterfactuals.

The derivability of conditionals $\alpha \Rightarrow \beta$ from a finite set $KB_{\Rightarrow}$ of conditionals by means of the rules above—resulting in the deductive consequence relations $\vdash_C$ and $\vdash_P$,

respectively—is defined just as usual, that is, analogously to the definition of derivability of formulas from formulas in classical propositional logic.

The following rules can be shown to be (meta-)derivable from the systems introduced above:

Lemma 8 *(Kraus, Lehmann, and Magidor 1990, pp. 179–80)*
The following rules are derivable in C (that is: from Reflexivity+Left Equivalence+Right Weakening+Cautious Cut+Cautious Monotonicity):

1. $\dfrac{\alpha \Rightarrow \beta, \alpha \Rightarrow \gamma}{\alpha \Rightarrow (\beta \wedge \gamma)}$ *(And)*
2. $\dfrac{\alpha \Rightarrow \beta, \beta \Rightarrow \alpha, \alpha \Rightarrow \gamma}{\beta \Rightarrow \gamma}$ *(Equivalence)*
3. $\dfrac{\alpha \Rightarrow (\beta \rightarrow \gamma), \alpha \Rightarrow \beta}{\alpha \Rightarrow \gamma}$ *(Modus Ponens in the Consequent)*
4. $\dfrac{\alpha \vdash \beta}{\alpha \Rightarrow \beta}$ *(Supra-Classicality)*

Lemma 9 *(Kraus, Lehmann, and Magidor 1990, p. 191)*
The following rules are derivable in P (that is: from Reflexivity+Left Equivalence+Right Weakening+Cautious Cut+Cautious Monotonicity+Disjunction; we label the derivable rules in the same way as Kraus, Lehmann, and Magidor 1990):

1. $\dfrac{\alpha \wedge \beta \Rightarrow \gamma}{\alpha \Rightarrow (\beta \rightarrow \gamma)}$ *(S)*
2. $\dfrac{\alpha \wedge \beta \Rightarrow \gamma, \alpha \wedge \neg\beta \Rightarrow \gamma}{\alpha \Rightarrow \gamma}$ *(D)*

Finally, we can relate the semantic systems of the previous subsection to the system of rules specified above by means of a soundness and completeness theorem (see Leitgeb 2004, chapter 10, for the theorem, and for the proofs in the relevant parts of the literature on which the theorem is based):

Theorem 10 *(Soundness and Completeness of P)*
Let $KB_{\Rightarrow} \subseteq \mathcal{L}_{\Rightarrow}$, $\alpha \Rightarrow \beta \in \mathcal{L}_{\Rightarrow}$; then each of the claims in theorem 7 is equivalent to:

$$KB_{\Rightarrow} \vdash_P \alpha \Rightarrow \beta$$

That is: the system P is sound and complete with respect to the continuity semantics, the sequence semantics, the uncertainty semantics, the order of magnitude semantics, the majority semantics, and the infinitesimal semantics for high probability conditionals.

In contrast, *none* of the following rules are (meta-)derivable in P nor are they valid with respect to any of the semantics of the last subsection, even though their counterparts for material conditionals are of course valid:

- $\dfrac{\alpha \Rightarrow \beta}{\neg\beta \Rightarrow \neg\alpha}$ (Contraposition)
- $\dfrac{\alpha \Rightarrow \beta,\ \beta \Rightarrow \gamma}{\alpha \Rightarrow \gamma}$ (Transitivity)
- $\dfrac{\alpha \Rightarrow \gamma}{\alpha \wedge \beta \Rightarrow \gamma}$ (Monotonicity; Strengthening of the Antecedent)

As Bennett (2003) argues in his chapter 9 (and as had been argued before by, e.g., Adams 1975 and Edgington 1995), none of these rules of inference is particularly plausible for the indicative if-then in natural language. Accordingly, in nonmonotonic reasoning all of these rules are normally given up as applying to default conditionals or (if reformulated accordingly) nonmonotonic consequence relations. However, we have already seen weakenings of these three rules to be contained in system P: in particular, Cautious Cut may be regarded as a weakening of Transitivity, and Cautious Monotonicity is clearly a weakened version of Monotonicity. (See Johnson and Parikh 2008 for an argument that, in a sense explained in their paper, the monotonicity rule is nevertheless 'almost valid' for probabilistic conditionals.)

The exchange of ideas between logic and probability theory, and the systematic study of jointly logical and probabilistic systems, has had a favourable effect on both areas in the past. It may have an even more favourable effect on the two areas in the future.

Acknowledgements

We are grateful to Stanislav Speranski, Alan Hájek, John Cusbert, and Edward Elliott for comments on a previous draft of this chapter. Work on this paper was supported generously by the Alexander von Humboldt Foundation.

References

Abadi, M. and Halpern, J. Y. (1994) Decidability and Expressiveness for First-Order Logics of Probability. *Information and Computation*. 112. pp. 1–36.

Adams, E. W. (1966) Probability and the Logic of Conditionals. In Hintikka, J. and P. Suppes (eds.) *Aspects of Inductive Logic*. pp. 265–316. Amsterdam: North-Holland.

Adams, E. W. (1974) The Logic of 'Almost All'. *Journal of Philosophical Logic*. 3. pp. 3–17.

Adams, E. W. (1975) *The Logic of Conditionals*. Dordrecht: Reidel.

Adams, E. W. (1986) On the Logic of High Probability. *Journal of Philosophical Logic*. 15. pp. 255–79.

Adams, E. W. (1996) Four Probability-Preserving Properties of Inferences. *Journal of Philosophical Logic*. 25. pp. 1–24.

Adams, E. W. (1998) *A Primer of Probability Logic*. Stanford, CA: CSLI Lecture Notes.

Adams, E. W. and Levine, H. P. (1975) On the Uncertainties Transmitted from Premisses to Conclusions in Deductive Inferences. *Synthese*. 30. pp. 429–60.

Arló-Costa, H. (2005) Non-Adjunctive Inference and Classical Modalities. *Journal of Philosophical Logic*. 34. pp. 581–605.

Arló-Costa, H. and Parikh, R. (2005) Conditional Probability and Defeasible Inference. *Journal of Philosophical Logic*. 34. pp. 97–119.

Bacchus, F. (1990a) On Probability Distributions Over Possible Worlds. In *Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence, UAI' 1988*. pp. 217–26. Amsterdam: North-Holland.

Bacchus, F. (1990b) *Representing and Reasoning with Probabilistic Knowledge*. Cambridge, MA: MIT Press.

Bacchus, F., Grove, A. J., Halpern, J. Y., and Koller, D. (1996) From Statistical Knowledge Bases to Degrees of Belief. *Artificial Intelligence*. 87. pp. 75–143.

Baltag, A. and Smets, S. (2008) Probabilistic Dynamic Belief Revision. *Synthese*. 165. pp. 179–202.

Bamber, D. (2000) Entailment with Near Surety of Scaled Assertions of High Conditional Probability. *Journal of Philosophical Logic*. 29. pp. 1–74.

Benferhat, S., Dubois, D., and Prade, H. (1997) Possibilistic and Standard Probabilistic Semantics of Conditional Knowledge. *Journal of Logic and Computation*. 9. pp. 873–95.

Benferhat, S., Saffiotti, A., and Smets, P. (2000) Belief Functions and Default Reasoning, *Artificial Intelligence*. 122. pp. 1–69.

Bennett, J. (2003) *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.

Biazzo, V., Gilio, A., Lukasiewicz, T., and Sanfilippo, G. (2002) Probabilistic Logic under Coherence, Model-Theoretic Probabilistic Logic, and Default Reasoning in System P. *Journal of Applied Non-Classical Logics*. 12. pp. 189–213.

Boole, G. (1854) *An Investigation of The Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. London: Macmillan.

Buchak, L. (2014) Belief, Credence, and Norms. *Philosophical Studies*. 169. pp. 285–311.

Burgess, J. P. (1969) Probability Logic. *The Journal of Symbolic Logic*. 34. pp. 264–74.

Caie, M. (2013) Rational Probabilistic Incoherence. *Philosophical Review*. 122. pp. 527–75.

Campbell-Moore, C. (2015a) Rational Probabilistic Incoherence? A Reply to Michael Caie. *The Philosophical Review*. 124. 3. pp. 393–406.

Campbell-Moore, C. (2015b) How to Express Self-Referential Probability: A Kripkean Proposal. *Review of Symbolic Logic*. 8. 4. pp. 680–704.

Campbell-Moore, C. (2016) Self-Referential Probability. Ph.D. Thesis in Philosophy. Ludwig-Maximilians-University Munich.

Carnap, R. (1950) *Logical Foundations of Probability*. Chicago, IL: University of Chicago Press.

Chang, C. C. and Keisler, H. J. (1990) *Model Theory*. Amsterdam: North-Holland.

Christensen, D. (2004) *Putting Logic in Its Place*. Oxford: Clarendon Press.

Christiano, P., Yudkowsky, E., Herreshoff, M., and Barasz, M. (unpublished) Definability of Truth in Probabilistic Logic.

Cross, C. B. (1993) From Worlds to Probabilities: A Probabilistic Semantics for Modal Logic. *Journal of Philosophical Logic*. 22. pp. 169–92.

Edgington, D. (1995) On Conditionals. *Mind*. 104. pp. 235–329.

Edgington, D. (1996) Vagueness by Degrees. In Keefe, R. and Smith, P. (eds.) *Vagueness: A Reader*. pp. 617–30. Cambridge, MA: MIT Press.

Fagin, R. (1976) Probabilities on Finite Models. *Journal of Symbolic Logic*. 41. pp. 50–8.

Fagin, R. and Halpern, J. Y. (1994) Reasoning about Knowledge and Probability. *Journal of the ACM*. 41. pp. 340–67.

Fagin, R., Halpern, J. Y., and Megiddo, N. (1990) A Logic for Reasoning About Probabilities. *Information and Computation*. 87. pp. 78–128.

Fenstad, J. E. (1967) Representations of Probabilities Defined on First Order Languages. In Crossley, J. N. (ed.) *Sets, Models and Recursion Theory*. pp. 156–172. Amsterdam: North-Holland.

Field, H. (1977) Logic, Meaning, and Conceptual Role. *The Journal of Philosophy*. 74. pp. 379–409.

Field, H. (2009) What is the Normative Role of Logic? *Proceedings of the Aristotelian Society*. Supplementary Volume LXXXIII. pp. 251–68.

Foley, R. (1993) *Working Without a Net*. Oxford: Oxford University Press.

Frisch, A. M. and Haddawy, P. (1988) Probability as a Modal Operator. In *Proceedings of the 4th Workshop on Uncertainty in AI*. pp. 109–18. Minneapolis, MN, Cornallis, OR: AUAI.

Frisch, A. M. and Haddawy, P. (1994) Anytime Deduction for Probabilistic Logic. *Artificial Intelligence*. 69. pp. 93–122.

Fuhrmann, A. (1997) *An Essay on Contraction*. Stanford, CA: CSLI Publications.

Gabbay, D. M. (1984) Theoretical Foundations for Non-Monotonic Reasoning in Expert Systems. In Apt, K. R. (ed.). *Logics and Models of Concurrent Systems*. pp. 439–58. Berlin: Springer.

Gabbay, D. M., Hogger, C. J., and Robinson, J. A. (eds.) (1994) *Handbook of Logic in Artificial Intelligence and Logic Programming*. 3. pp. 35–110. Oxford: Clarendon Press.

Gaifman, H. (1964) Concerning Measures in First Order Calculi. *Israel Journal of Mathematics*. 2. pp. 1–18.

Gaifman, H. (1986) A Theory of Higher Order Probabilities. In *Proceedings of the Conference on Theoretical Aspects of Reasoning about Knowledge*. pp. 275–92. Monterey, California.

Gaifman, H. and Snir, M. (1982) Probabilities Over Rich Languages, Testing and Randomness. *The Journal of Symbolic Logic*. 47. pp. 495–548.

Gärdenfors, P. (1975) Qualitative Probability as an Intensional Logic. *Journal of Philosophical Logic*. 4. pp. 171–85.

Gärdenfors, P. and Makinson, D. (1994) Nonmonotonic Inference Based on Expectations. *Artificial Intelligence*. 65. pp. 197–245.

Goldszmidt, M. and Pearl, J. (1996) Qualitative Probabilities for Default Reasoning, Belief Revision, and Causal Modeling. *Artificial Intelligence*. 84. pp. 57–112.

Haenni, R. (2005) Unifying Logical and Probabilistic Reasoning. In Godo, L. (ed.) *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*. Lecture Notes in Artificial Intelligence. Vol. 3571. pp. 788–99. Berlin: Springer.

Hailperin, T. (1937) Foundations of Probability in Mathematical Logic. *Philosophy of Science*. 4. pp. 125–50.

Hailperin, T. (1984) Probability Logic. *Notre Dame Journal of Formal Logic*. 25. pp. 198–212.

Hailperin, T. (1996) *Sentential Probability Logic*. Bethlehem, PA: Lehigh University Press.

Hailperin, T. (2000) Probability Semantics for Quantifier Logic. *Journal of Philosophical Logic*. 29. pp. 207–39.

Hájek, A. (1989) Probabilities of Conditionals–Revisited. *Journal of Philosophical Logic*. 18. pp. 423–28.

Halpern, J. Y. (1990) An Analysis of First-Order Logics of Probability. *Artificial Intelligence*. 46. pp. 311–50.

Halpern, J. Y. (1991) The Relationship between Knowledge, Belief, and Certainty. *Annals of Mathematics and Artificial Intelligence*. 4. pp. 301–22.

Halpern, J. Y. (2001) Lexicographic Probability, Conditional Probability, and Nonstandard Probability. In *Proceedings of the Eighth Conference on Theoretical Aspects of Rationality and Knowledge*. pp. 17–30. Ithaca, NY: Morgan Kaufmann.

Halpern, J. Y. (2003) *Reasoning About Uncertainty*. Cambridge, MA: MIT Press.

Halpern, J. Y. and Rabin, M. O. (1987) A Logic to Reason about Likelihood. *Artificial Intelligence*. 32. pp. 379–405.

Hamblin, C. L. (1959) The Modal 'Probably'. *Mind*. 68. pp. 234–40.

Hawthorne, J. (1996) On the Logic of Nonmonotonic Conditionals and Conditional Probabilities. *Journal of Philosophical Logic*. 25. pp. 185–218.

Hawthorne, J. (2007) Nonmonotonic Conditionals that Behave Like Conditional Probabilities Above a Threshold. *Journal of Applied Logic*. 5. pp. 625–37.

Hawthorne, J. and Makinson, D. (2007) The Quantitative/Qualitative Watershed for Rules of Uncertain Inference. *Studia Logica*. 86. pp. 247–97.

Heifetz, A. and Mongin, P. (2001) Probability Logic for Type-Spaces. *Games and Economic Behavior*. 35. pp. 31–53.

Hempel, C. G. (1962) Deductive-Nomological vs Statistical Explanation. In Feigl, H. and Maxwell, G. (eds.) *Minnesota Studies in the Philosophy of Science*. 3. pp. 98–169. Minneapolis, MN: University of Minnesota Press.

Hilpinen, R. (1968) Rules of Acceptance and Inductive Logic. *Acta Philosophical Fennica*. 22. Amsterdam: North-Holland.

Hintikka, J. and P. Suppes (eds.) (1966) *Aspects of Inductive Logic*. Amsterdam: North-Holland.

Hoover, D. N. (1978) Probability Logic. *Annals of Mathematical Logic*. 14. pp. 287–313.

Howson, C. (2003) Probability and Logic. *Journal of Applied Logic*. 1. pp. 151–65.

Huber, F. and Schmidt-Petri, C. (eds.) (2009) *Degrees of Belief*. Synthese Library. 342. Berlin: Springer.

Johnson, M. and Parikh, R. (2008) Probabilistic Conditionals are Almost Monotonic. *Review of Symbolic Logic*. 1. pp. 73–80.

Keisler, H. J. (1985) Probability Quantifiers. In Barwise, J. and Feferman, S. (eds.) *Model-Theoretic Logics*. pp. 509–56. New York, NY: Springer.

Kooi, B. P. (2003) Probabilistic Dynamic Epistemic Logic. *Journal of Logic, Language and Information*. 12. pp. 381–408.

Kraus, S., Lehmann, D., and Magidor, M. (1990) Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*. 44. pp. 167–207.

Kyburg, H. Jr. (1961) *Probability and the Logic of Rational Belief*. Middletown, CT: Wesleyan University Press.

Lando, T. (2010) Completeness of S4 for the Lebesgue Measure Algebra. *Journal of Philosophical Logic*. 41. pp. 287–316.

Leblanc, H. (1979) Probabilistic Semantics for First-Order Logic. *Zeitschrift für mathematische Logik und Grundlagen der Mathematik*. 25. pp. 497–509.

Leblanc, H. (1983) Alternatives to Standard First-Order Semantics. In Gabbay, D. and Guenthner, F. (eds.) *Handbook of Philosophical Logic*. Vol. 1. pp. 189–274. Dordrecht: Reidel.

Lehmann, D. and Magidor, M. (1992) What Does a Conditional Knowledge Base Entail? *Artificial Intelligence*. 55. pp. 1–60.

Leitgeb, H. (2004) *Inference on the Low Level: An Investigation into Deduction, Nonmonotonic Reasoning, and the Philosophy of Cognition*. Dordrecht: Kluwer.

Leitgeb, H. (2012a) A Probabilistic Semantics for Counterfactuals. Part A. *Review of Symbolic Logic*. 5. pp. 16–84.
Leitgeb, H. (2012b) A Probabilistic Semantics for Counterfactuals. Part B. *Review of Symbolic Logic*. 5. pp. 85–121.
Leitgeb, H. (2012c) From Type-Free Truth to Type-Free Probability. In Restall, G. and Russell, G. (eds.) *New Waves in Philosophical Logic*. pp. 84–93. New York, NY: Palgrave Macmillan.
Leitgeb, H. (2014) The Stability Theory of Belief. *The Philosophical Review*. 123. pp. 131–71.
Levi, I. (1967) *Gambling with the Truth: An Essay on Induction and the Aims of Science*. Cambridge, MA: MIT Press.
Lewis, D. (1973) *Counterfactuals*. Oxford: Basil Blackwell.
Lewis, D. (1980) A Subjectivist's Guide to Objective Chance. In Jeffrey, R.C. (ed.), *Studies in Inductive Logic and Probability*, vol. 2. pp. 263–93. Berkeley: University of California Press.
Lewis, D. K. (1976) Probabilities of Conditionals and Conditional Probabilities. In *Philosophical Papers*. Vol. II. pp. 133–56. Oxford: Oxford University Press.
Lewis, D. K. (1986) A Subjectivist's Guide to Objective Chance. In *Philosophical Papers*. Vol. II. pp. 83–132. Oxford: Oxford University Press.
Lin, H. and Kelly, K. T. (2012) Propositional Reasoning that Tracks Probabilistic Reasoning. *Journal of Philosophical Logic*. 41. pp. 957–81.
Maher, P. (1993) *Betting on Theories*. Cambridge: Cambridge University Press.
Makinson, D. (2011) Conditional Probability in the Light of Qualitative Belief Change. *Journal of Philosophical Logic*. 40. pp. 121–53.
Makinson, D. (2012) Logical Questions behind the Lottery and Preface Paradoxes: Lossy Rules for Uncertain Inference. *Synthese*. 186. pp. 511–29.
McGee, V. (1989) Conditional Probabilities and Compounds of Conditionals. *The Philosophical Review*. 98. pp. 485–541.
Morgan, C. (1982) Simple Probabilistic Semantics for Modal Logic. *Journal of Philosophical Logic*. 11. pp. 443–58.
Nilsson, N. (1986) Probabilistic Logic. *Artificial Intelligence*. 28. pp. 71–87.
Paris, J. (2011) Pure Inductive Logic. In Horsten, L. and Pettigrew, R. (eds.). *The Continuum Companion to Philosophical Logic*. pp. 428–49. London: Continuum.
Paris, J. and Simmonds, R. (2009) O Is Not Enough. *Review of Symbolic Logic*. 2. pp. 298–309.
Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.
Pfeifer, N. and Kleiter, G. D. (2005) Coherence and Nonmonotonicity in Human Reasoning. *Synthese*. 146. pp. 93–109.
Pfeifer, N. and Kleiter G. D. (2010) The Conditional in Mental Probability Logic. In Oaksford, M. and Chater, N. (eds.) *Cognition and Conditionals: Probability and Logic in Human Thought*. pp. 153–73. Oxford: Oxford University Press.
Popper, K. R. (1955) Two Autonomous Axiom Systems for the Calculus of Probabilities. *British Journal for the Philosophy of Science*. 6. pp. 51–7.
Ramsey, F. P. (1926) Truth and Probability. In Ramsey, F. P. (ed.) *The Foundations of Mathematics and other Logical Essays*. pp. 156–98. London: Kegan Paul.
Rescher, N. (1962) A Probabilistic Approach to Modal Logic. *Acta Philosophica Fennica*. 16. pp. 215–26.
Richardson, M. and Domingos, P. (2006) Markov Logic Networks. *Machine Learning*. 62. pp. 107–36.

Roeper, P. and Leblanc, H. (1999) *Probability Theory and Probability Semantics*. Toronto: University of Toronto Press.
Ross, J. and Schroeder, M. (2014) Belief, Credence, and Pragmatic Encroachment. *Philosophy and Phenomenological Research*. 88. 2. pp. 259–88.
Schurz, G. (1997) Probabilistic Default Logic Based on Irrelevance and Relevance Assumptions. In Gabbay, D. M. et al. (eds.) *Qualitative and Quantitative Practical Reasoning*. pp. 536–553. Berlin: Springer.
Schurz, G. (1998) Probabilistic Semantics for Delgrande's Conditional Logic and a Counterexample to his Default Logic. *Artificial Intelligence*. 102. pp. 81–95.
Scott, D. and Krauss, P. (1966) Assigning Probabilities to Logical Formulas. In Hintikka, J. and P. Suppes (eds.) *Aspects of Inductive Logic*. pp. 219–264. Amsterdam: North-Holland.
Segerberg, K. (1971) Qualitative Probability in a Modal Setting. In Fenstad, J. E. (ed.) *Proceedings of the 2nd Scandinavian Logic Symposium*. pp. 341–52. Amsterdam: North-Holland.
Snow, P. (1999) Diverse Confidence Levels in a Probabilistic Semantics for Conditional Logics. *Artificial Intelligence*. 113. pp. 269–79.
Speranski, S. O. (2013) Complexity for Probability Logic with Quantifiers over Propositions. *Journal of Logic and Computation*. 23. pp. 1035–55.
Spohn, W. (1988) Ordinal Conditional Functions: A Dynamic Theory of Epistemic States. In Harper, W. L. and Skyrms, B. (eds.) *Causation in Decision, Belief Change, and Statistics*. 2. pp. 105–34. Dordrecht: Reidel.
Spohn, W. (2012) *The Laws of Belief: Ranking Theory and Its Philosophical Applications*. Oxford: Oxford University Press.
Stalnaker, R. C. (1968) A Theory of Conditionals. In Rescher, N. (ed.) *Studies in Logical Theory*. pp. 98–112. Blackwell.
Sturgeon, S. (2008) Reason and the Grain of Belief. *Noûs*. 42. pp. 139–65.
Suppes, P. (1966) Probabilistic Inference and the Concept of Total Evidence. In Hintikka, J. and P. Suppes (eds.) *Aspects of Inductive Logic*. pp. 49–65. Amsterdam: North-Holland.
Swain, M. (ed.) (1970) *Induction, Acceptance and Rational Belief*. Dordrecht: Reidel.
Terwijn, S. A. (2005) Probabilistic Logic and Induction. *Journal of Logic and Computation*. 15. pp. 507–15.
van Benthem, J., Gerbrandy, J., and Kooi, B. (2009) Dynamic Update with Probabilities. *Studia Logica*. 93. pp. 67–96.
van Fraassen, B. (1981) Probabilistic Semantics Objectified: I. Postulates and Logics. *Journal of Philosophical Logic*. 10. pp. 371–94.
van Fraassen, B. (1995) Belief and the Problem of Ulysses and the Sirens. *Philosophical Studies*. 77. pp. 7–37.
Wedgwood, R. (2012) Outright Belief. *Dialectica*. 66. 3. pp. 309–29.
Wheeler, G. (2007) A Review of the Lottery Paradox. In Harper, W. L. and Wheeler, G. (eds.). *Probability and Inference: Essays in Honor of Henry E. Kyburg Jr.* pp. 1–31. London: King's College Publications.
Yalcin, S. (2010) Probability Operators. *Philosophy Compass*. 5. pp. 916–37.

CHAPTER 33

PROBABILITY IN ETHICS

DAVID MCCARTHY

Ethics is mainly about what we ought to do, and about when one situation is better than another. But facing uncertainty about the consequences of our actions, and about how situations will evolve, is an all-pervasive feature of our condition. Should this not be a central topic in ethical theory?

Probability is by far the best-known tool for thinking about uncertainty, a well-known aphorism telling us that it is the very guide to life. But despite important exceptions, it is easy to get the impression that mainstream moral philosophy has not been much concerned with probability.

This reflects what seems to be a natural division of labour. The most fundamental questions for ethical theory seem to arise in the absence of uncertainty. For example, it seems hard to believe that the questions of whether it is better to give priority to the worse off, and of whether we ought to favour our nearest and dearest, have anything to do with uncertainty. Many influential discussions of these topics never mention uncertainty.

Of course, once answers to these fundamental questions are in, we can try to extend them to cases involving uncertainty. But ethical theorists may seem well advised to hand this task over to others, given how mathematical the various disciplines concerned with probability have become. Technically and philosophically interesting as it may be, the extension of central ethical ideas to problems involving probability seems to be outside the main business of ethical theory.

This chapter will argue for the opposite view. The major ethical problems to do with probability involve very little mathematics to appreciate; many topics which do not seem to have anything to do with probability are arguably all about probability; and thinking about various problems to do with probability can help us solve analogous problems which do not involve probability, sometimes even revealing that popular positions about such problems are incoherent.

Almost every topic discussed here could easily be given its own survey article, and an adequate bibliography would exceed the space allotted for the whole chapter. Positive positions are often argued for sketchily, many important positions on each topic are neglected, and some major topics are not discussed at all.

Instead, the goal is to offer enough breadth to illustrate some ways in which questions about probability run systematically throughout ethical theory, while in places going into

enough depth to articulate some surprising and potentially important applications. In brief, what follows is much less a survey of or an argument for particular positions than a plea for ethical theory to take probability more seriously.

I said that ethics is largely about what we ought to do, and when one situation is better than another. Some say that rationality is about these things as well. Given that theories of rationality in the face of uncertainty are highly developed, it might be thought that an appeal to these theories of rationality straightforwardly solves ethical problems about probability.

This line of thought is importantly mistaken. First, Hume famously claimed that it is not irrational for an agent to prefer the destruction of the whole world to the scratching of his finger. Nor would it be irrational for the agent to bring about the destruction to avoid the scratching. But the destruction is neither better than the scratching, nor better for the agent. And the agent surely ought not to bring about the destruction. On at least one widely-held view, therefore, ethics and rationality are not about the same things.

Secondly, it is undeniable that contemporary theories of rationality are an indispensable resource for thinking about ethics and probability. However, whether and how to apply these theories to ethics is far from straightforward, and will be one of the principal concerns of this chapter. Furthermore, in my view, at least, appeals to rationality are almost always epiphenomenal. For example, suppose we have a convincing argument for the claim that rational preferences have such and such a structure. We could then try to claim that an evaluative relation like betterness has to have that structure on the grounds that a rational agent can surely prefer what's better to what's worse. However, it is almost always less committal and more direct just to modify the original argument to make it apply directly to the structure of the evaluative relation. Claims about rationality often have historical priority over parallel claims about ethics, but I believe they do not have any kind of important conceptual priority.

The chapter starts with four sections which discuss which probabilities are relevant to ethics, establish terminology, and rehearse expected utility theory. It then turns to the evaluative question of when one situation is better than another, focusing on the question of when one distribution of goods is better than another. Sections 33.5 and 33.6 discuss popular but I think inadequate approaches to this question. These serve as a backdrop to a hugely important theorem due to Harsanyi (1955) introduced in section 33.7. Sections 33.8 to 33.16 discuss such things as the relationship between Harsanyi's theorem and utilitarianism; criticisms of Harsanyi's premises and the relationship of these criticisms to other distributive views such as egalitarianism, the priority view, and concerns with fairness; the extension of Harsanyi's theorem to problems of population size; incommensurability; continuity; non-expected utility theory; evaluative measurement; and the question of what Harsanyi's theorem really shows about aggregation. These sections also list various open problems and directions for further work. All of these topics have to do with probability.

One of the benefits of thinking about Harsanyi's theorem is the way it helps us organize our thinking about all sorts of fundamental evaluative questions. Section 33.17 will suggest that thinking about decision theory can have the same value in thinking about fundamental normative questions, questions about what we ought to do. With particular focus on probability, the remaining sections illustrate by discussing what are arguably the three most important kinds of normative theories: act consequentialism, rule consequentialism or contractualism, and deontology (these will be defined in section 33.3).

The discussion aims to be self-contained. For those with a background in ethics who would like to know more about how probability is involved, the chapter keeps technicalities

to a minimum. But the topic just cannot be addressed without a certain amount of rigor, and passing acquaintance with expected utility theory and decision theory will be helpful, though not strictly necessary. For those who know about probability and would like to see how it applies to ethics, the chapter gives brief guides to the relevant ethical debates. Such readers will recognize occasional allusions to relatively sophisticated ideas to do with probability. For one thing is clear: the questions about probability which ethics raises are profound, and are surely best addressed by combining expertise.

33.1 Probabilities

One difficulty in thinking about probability in ethics is assessing when ethicists need to be involved. Suppose we are told that some action will benefit many but involves a small probability of harming a few. We might think it the job of epistemologists, metaphysicians or philosophers of science to tell us what kind of judgment 'the probability is small' expresses, what laws probabilities obey, and what makes such a judgment correct. Ethicists need only ask whether we ought to perform the action given that the probability of harm is small, and need not be involved any earlier. However, the division of labour is unlikely to be so neat.

There are many conceptions of probability (see e.g. Hájek, 2012, for a survey). This raises the question of which conception is most relevant to ethics, or whether different conceptions are appropriate in different ethical contexts. One of the most basic distinctions is between subjective and objective conceptions of probability, and this distinction will enable us to illustrate many of the issues.

The best-known subjective conception claims that the preferences of an ideally rational agent between uncertain prospects must satisfy various structural conditions (Ramsey, 1931; Savage, 1954). Suppose the agent also has a rich set of preferences. Then Ramsey and Savage showed that there exists a unique function on events satisfying the usual probability axioms (call it her subjective probability function) and a function on outcomes (her utility function) such that: the agent weakly prefers one prospect to another if and only if the former has at least as great expected utility, as calculated by those functions.

Perhaps the most prominent objective conception of probability in the contemporary debate is the best-system analysis pioneered by Lewis. The original best-system analysis of the laws of nature of Lewis (1973) says that the laws are the theorems of the best systematization of the world: the true theory which does best in terms of simplicity and strength (or informativeness). To allow probabilistic laws in, Lewis (1994) introduced the idea of fit. The more likely the actual world is by the lights of the theory, the better the fit of that theory. Theories are now judged according to how well they do in terms of simplicity, strength, and fit. If some of the laws of the best theory are probabilistic, those are what determine the objective probabilities.

Suppose we have to choose between subjective and objective conceptions for use in ethics, understood along the lines just sketched.[1] Which conception should it be?

[1] Neither the Ramsey-Savage story about subjective probabilities nor Lewis's version of best-system analysis has a hegemony. For surveys of alternative views about subjective probability, see Gilboa (2009), and for alternative best-system analyses, see Schwarz in this volume (2016).

Perhaps it depends on context: for example, subjective probabilities may be appropriate for agent-evaluation (blame, responsibility etc.), but inappropriate in other contexts. But let us fix the context by focussing on the most basic normative question of what we ought to do.

Each conception has features we might find appealing. Objective probabilities seem in some important sense to trump subjective probabilities. This is reflected in the popular view that when an agent has beliefs about objective probabilities, rationality requires her to conform her subjective probabilities to those beliefs. This is the basic idea behind the so-called principle principal of Lewis (1980). But if objective probabilities do indeed trump subjective probabilities, it may seem that what we ought to do depends on the objective probabilities, not our subjective probabilities.

On the other hand, objective probabilities may be disappointingly sparse or epistemically inaccessible. For example, best-systems analyses may make good sense of the objective probability of radium atoms decaying or coins landing on heads. But it is much less clear what best-systems analyses have to say about the objective probability of events like a run on a particular bank next year, one-off macro events involving chaotic systems. Such events may fail to have reasonably determinate objective probabilities (compare Hoefer, 2007), and even if they do, the epistemology may be too difficult for the objective probabilities to be usefully action guiding.

So perhaps we should instead say that what we ought to do depends at least in part on our subjective probabilities. One option is to use subjective probabilities exclusively; another is to use objective probabilities where available, and subjective probabilities to fill in the gaps. But every view which makes significant use of subjective probabilities faces at least two major problems.

First, the Ramsey-Savage story about subjective probabilities is a chapter in the Humean story about rationality. But just as the Humean story refuses to condemn the preference for the destruction of the whole world over the scratching of a finger, the Ramsey-Savage story does not condemn subjective probabilities which, to most people, are just as crazy. For example, provided her preferences are appropriately structured, there is nothing in the Ramsey-Savage story to condemn someone who thinks it highly likely the world will come to an end before teatime. Such subjective probabilities will seem to many too irrational to have any bearing on what we ought to do. But it is a major challenge to articulate a principled account of which subjective probabilities should be excluded.

Secondly, as soon as we allow in subjective probabilities, we face questions of *whose* and *how*. Whose subjective probabilities count in determining whether an agent ought to perform some action – the agent's, those of her potential victims or beneficiaries, everyone's? If the subjective probabilities of at least two people are relevant, how should they be used? At least if we switch to the problem of evaluating the uncertain prospects which actions result in, this is a long-standing problem in welfare economics. The so-called *ex post* approach recommends first aggregating the separate subjective probability functions into a single social probability function, then using this social probability function to evaluate uncertain prospects. The *ex ante* approach gives the separate subjective probability functions a direct evaluative role, at least in a special case. Just to give one version, *ex ante Pareto* says: if for each individual i, an uncertain prospect P is better for i than another uncertain prospect P' relative to i's own subjective probability function, then P is better than P'. Both the *ex post* and *ex ante* approaches look appealing, but they are extremely difficult to combine consistently. For example, given weak assumptions, there will be prospects P and P' such

that *ex ante* Pareto has the apparent pathology of implying that P is better than P' despite the fact that P' is guaranteed to produce a better outcome than P. But *ex post* approaches will adopt principles which from the outset say that in such cases, P' is better than P.[2]

Now it is not my goal to try to answer any of the large questions raised in this section. My claim is rather that they are questions with which ethicists must engage, and that one's answers to these questions may depend on one's more general ethical views. To illustrate, suppose one sees ethics as being primarily about coordinating action to achieve good outcomes, and one is prepared to tolerate a significant amount of indeterminacy in one's normative theory. Then one may be tempted to claim that the probabilities which are relevant to ethics are the objective probabilities alone. By contrast, suppose one instead sees ethics as being about trying to achieve some sort of fair compromise between agents with diverse beliefs and goals. Then it may seem tempting to allow in subjective probabilities no matter how irrational, and to follow the *ex ante* approach. On this picture, individual autonomy is central, and it may seem more important to respect the notion of unanimity built into *ex ante* Pareto than to try to avoid the apparent pathology which comes with it. There are, of course, many other options, but the important point is that which probabilities are relevant to ethics, and how, is itself a fundamental ethical question.

33.2 Outcomes

Some writers, however, think that probabilities are never relevant to what we ought to do. A parallel view applies to the question of when one uncertain prospect is better than another. Jackson (1991) illustrates with the following. A doctor has to choose between three treatments for a patient with a minor complaint. Drug A would partially cure the complaint. One of drugs B and C would completely cure the patient while the other would kill him, but the doctor cannot tell which is which.

The obvious view, as Jackson notes, is that the doctor ought to give the patient drug A. This verdict would be delivered by any broadly decision-theoretic account. Along similar lines, the prospect associated with giving the patient drug A is better than the prospects associated with drugs B and C. Call any view which assesses actions and prospects involving uncertainty along broadly decision-theoretic lines *probability-based*.

But there is a different view: if drug B would cure the patient, the doctor ought to give the patient drug B; similarly for drug C. Likewise, if drug B would cure, the prospect associated with giving drug B is better than the other prospects. Call such views, positions which assess actions and prospects in terms of what their consequences would be, *outcome-based*.

As the drug example shows, an objection to outcome-based views is that they make the truth about what we ought to do too epistemically inaccessible, or provide poor guides to action. But there are at least two interesting arguments for outcome-based views.

First, transposing an argument due to Thomson (1986) to the present example, suppose the pharmacist walks in and knowing full-well that drug B would cure the patient, says to

[2] The large literature on this topic is rather technical, but Broome (1991, ch. 7) provides a good introduction and philosophical discussion. Mongin (1995) contains a very general set of results.

the doctor: "You ought to use drug *B*". The pharmacist seems right. But doesn't that imply an outcome-based view?

In response, consider the case where the pharmacist says: "Drug *B* would cure. So you ought to use drug *B*". By the time the pharmacist has finished the first sentence, the doctor has new evidence, and should upgrade her probabilities accordingly. There is then no clash between a probability-based view and the truth of the pharmacist's second sentence. Likewise, I think that in the actual case something like "Drug *B* would cure" is implied when the pharmacist just says: "You ought to use drug *B*". What is implied by the pharmacist's normative assertion impacts upon the probabilities the doctor should have, making the literal construal of the normative assertion true (McCarthy, 1998).

Secondly, advocates of probability-based views have to say which probability functions are relevant to what we ought to do. But there are many candidates, e.g. the probabilities of this agent or that agent, at this time or that time. Jackson concludes that we have to recognize the existence of "an annoying profusion ... of a whole range of oughts" (Jackson, 1991, p. 471).

But this seems dissatisfying. When we ask ourselves or others what we ought to do, we don't want to learn that some oughts recommend this while others recommend that. We want to know what we ought to do *fullstop*. But if there is only one ought, we need to privilege one probability function. The function of an omniscient agent may seem to be the only distinguished choice, so we end up with an outcome-based view.

In response, just because it is not obvious which probability function is privileged, it does not follow that no function (or reasonably narrow class of functions) is privileged. In the previous section we saw that if we adopt a probability-based view, a variety of fairly fundamental ethical factors and disputes bears upon the question of which probabilities are relevant to ethics. The complexity of this topic explains why it is not obvious which probability function is privileged, but the fact that the problem is complex hardly entails that some outcome-based view wins by default. Outcome-based views have to be assessed in terms of various ethical desiderata just as much as probability-based views do, and they do quite badly in terms of desiderata such as the idea that an ethical theory should be suitably action-guiding.

It is also worth noting that outcome-based views may result in large-scale indeterminacy. The drug example stipulated that various counterfactuals relating actions to outcomes are true. But an increasingly popular view claims that most counterfactuals are false (see e.g. Hájek, 2014). In particular, it will often be the case that for some potential action *A* there is no outcome *O* such that the counterfactual: "If *A* were performed, *O* would result", is true. On this view about counterfactuals, the facts on which outcome-based views have to call are much sparser than might have appeared, with the result that there is a lot more evaluative and normative indeterminacy on outcome-based views than we might have hoped. This may further undercut the appeal of outcome-based views.

In what follows, I will assume that some probability-based view is correct. But it is a major question which conception of probability is relevant to ethics, so ethicists need to be involved with questions about probability early on. In light of the difficulties of aggregating probability functions alluded to in the previous section, ethicists also need to be prepared for the possibility that the eventual input into ethics is going to be messier than a single probability function which satisfies the usual axioms.

33.3 Terminology

However, to simplify I henceforth assume that probabilities are supplied and satisfy the usual axioms. To reflect this I will often speak of *risk* rather than probability or uncertainty. A *lottery* over a nonempty set of world histories (past, present and future) assigns positive probabilities to finitely many of the histories with the probabilities all summing to one (these are sometimes known as lotteries with *finite support*). I will often write lotteries in the form $[p_1,h_1;\ldots;p_m,h_m]$ where the h_j's are the histories which could result from the lottery and the p_j's their probabilities.

The *betterness relation* holds between two lotteries just in case the first is at least as good as the second. An individual *i*'s *individual betterness relation* holds between two lotteries L_1 and L_2 just in case: *i* exists in every history which could result from the lotteries, and L_1 is at least as good for *i* as L_2. By identifying histories with lotteries in which the history gets probability one, and restricting the betterness and individual betterness relations to such lotteries, we obtain relations between histories. I will refer to these relations as *risk-free* versions of the originals. For example, the *risk-free betterness relation* holds between two histories just in case the first is at least as good as the second.

There are many views about when one history is better for someone than another, or in a more suggestive phrase, about what makes someone's life go best (Parfit, 1984, Appendix I). On one popular classification, the three main views are that having a good life is a matter of: (i) having good-quality experiences; (ii) satisfying one's preferences or desires; or (iii) attaining what are said to be objective goods, such as deep knowledge or close personal relationships. However, some philosophers think that when doing ethics, we should not be in the business of making fine-grained comparisons between different people's lives, but should make interpersonal comparisons only in terms of such things as the resources, freedoms, or opportunities people enjoy (see e.g. Rawls, 1982; Sen, 1985). Which of these views is correct will not matter in what follows, but it will be important that the discussion can accommodate any of them.

We will be talking a lot about the betterness relation. Not everyone thinks that this is a useful way of looking at ethics (see e.g. Foot, 1985; Thomson, 2001). But in response, talking about betterness can be seen as a harmless organizing tool (see e.g. Broome, 1991), and is popular enough for us to be able to cover many major positions. For example, *consequentialism* (on a probability-based interpretation) is the view that lotteries can be ranked in terms of betterness, and that betterness somehow determines normativity.[3] For example, *act consequentialism* says that we always ought to bring about the best available lottery, whereas *rule consequentialism* says that we always ought to act according to the rule such that, if everyone acted in accord with it (or on a different version, accepted it), the best available lottery would be realized. *Contractualism* tends to be framed not in terms of betterness, but in terms of an ideal social contract. However, when it comes to the assessment of different social contracts, contractualists are concerned with

[3] As far as I can see, there is no universally accepted account of consequentialism, so I am only trying to convey the rough idea rather than provide a precise definition. In addition, the way moral philosophers use the term 'consequentialism' should not be confused with an important decision-theoretic idea which also goes by the name of 'consequentialism' (see e.g. Hammond, 1998).

competing sets of principles or rules (see e.g. Scanlon, 1982), so at the concrete level of normative theorizing, it is often hard to tell the difference between contractualism and rule consequentialism. Finally, *deontology* is often characterized as the position that some acts are wrong even when they would have the best available consequences, such as killing one innocent person to prevent five innocent people from being killed.

33.4 Expected utility theory

This chapter expresses the view that whatever one ultimately makes of expected utility theory and decision theory, looking at basic evaluative and normative questions through the frameworks they provide is extremely useful. This section therefore provides a quick rehearsal, first of the terminology of expected utility theory, and then of its most basic result. It takes X to be some fixed nonempty set. In applications, X will usually be a set of histories, or more colloquially, outcomes.

A *preorder* on X is a binary relation R on X which is reflexive ($\forall x \in X$, xRx) and transitive ($\forall x,y,z \in X$, xRy & $yRz \implies xRz$). It is *complete* if for all $x,y \in X$, either xRy or yRx. It is *incomplete* just in case it is not complete. An *ordering* of X is just a complete preorder of X. If L and M are lotteries over X, then for all $\alpha \in (0,1)$, $\alpha L + (1-\alpha)M$ is the so-called *compound lottery* in which each member x of X has probability $\alpha p + (1-\alpha)q$ where p is x's probability under L and q is its probability under M. Suppose that $\succsim$ is an ordering on X. Then a real-valued function f is said to *represent* the ordering just in case: for every x and y in X, $x \succsim y$ if and only if $f(x) \geq f(y)$.

Suppose that $\succsim$ is a binary relation on lotteries over X. Here are the three expected utility axioms.

Ordering $\succsim$ is a complete preorder.

Strong Independence For all lotteries L, M and N, and $\alpha \in (0,1)$: $L \succsim M$ if and only if $\alpha L + (1-\alpha)N \succsim \alpha M + (1-\alpha)N$.

The rough idea of Strong Independence is that the "addition" of the same lottery N to either side of $L \succsim M$ should make no difference: the added N's will cancel out. Strong Independence is sometimes explained by imagining that the compound lotteries will be realized by first tossing a biased coin, where heads has a probability of α and tails a probability of $1-\alpha$, then running whichever lottery results. For example, suppose you strictly prefer L to M, and you now have to decide between $\alpha L + (1-\alpha)N$ and $\alpha M + (1-\alpha)N$. If the coin lands on tails, you will face N in either case, so in that scenario there is nothing to choose between the two compound lotteries. But if the coin lands on heads, you will face L or M, and will therefore prefer to have chosen $\alpha L + (1-\alpha)N$ to $\alpha M + (1-\alpha)N$. Since heads has a positive probability, you should therefore strictly prefer $\alpha L + (1-\alpha)N$ to $\alpha M + (1-\alpha)N$ prior to the coin being tossed. Or at least that is one of the typical ways of motivating Strong Independence. The example has focused on preference relations, but it can clearly be applied directly and without any discussion of rationality to a variety of evaluative comparatives, such as betterness and individual betterness relations.

Continuity For all lotteries L, M and M such that $L \succ M \succ N$,[4] there exist $\alpha, \beta \in (0,1)$ such that $M \succ \alpha L + (1-\alpha)N$ and $\beta L + (1-\beta)N \succ M$.

To illustrate, suppose you strictly prefer \$1000 to \$100, and strictly prefer \$100 to \$10. Then if your preferences are continuous, there will be some lottery which almost guarantees you \$1000 with a tiny chance of \$10 (one in a billion, say) which you will strictly prefer to getting \$100 for certain. And you will strictly prefer \$100 for certain to some lottery which almost guarantees you \$10 with a tiny chance of \$1000. As the example is meant to suggest, many people think that Continuity is a plausible requirement on various evaluative comparatives.

A binary relation $\succsim$ on lotteries over X satisfies the expected utility axioms just in case it satisfies Ordering, Strong Independence, and Continuity. Here is the most basic result of expected utility theory, due to von Neumann and Morgenstern (1944), but anticipated in a deeper way by Ramsey (1931).

Theorem 1 *(von Neumann and Morgenstern) Let X be a nonempty set, and $\succsim$ be a binary relation on lotteries on X which satisfies the expected utility axioms. Then there exists a real-valued function u on X such that*

1. *For all lotteries $L_1 = [p_1, x_1; \ldots; p_m, x_m]$ and $L_2 = [q_1, y_1; \ldots; q_n, y_n]$,*

$$L_1 \succsim L_2 \iff p_1 u(x_1) + \cdots + p_m u(x_m) \geq q_1 u(y_1) + \cdots + q_n u(y_n)$$

2. *Any function v satisfies (i) when substituted for u if and only if there exist real numbers $a > 0$ and b such that $v = au + b$.*

Roughly speaking, (i) says that there is a function u (often referred to as a "vNM utility function") such that $L_1 \succsim L_2$ if and only if the expected value of u associated with L_1 is at least as great as the expected value of u associated with L_1. The expected value of u associated with a lottery is obtained by applying u to each of the lottery's possible outcomes, weighting the result by the probability of those outcomes, then adding all those numbers up. In such circumstances, I will say that the ordering $\succsim$ is *represented by the expected value of u*. (ii) says that the function u is unique up to choice of zero and unit, or in fancier terminology, unique up to positive affine transformation. For an analogy, Fahrenheit and Centigrade measure temperature in essentially the same way, except that they use different zeros and units. Overall, the main message is that if an ordering of lotteries satisfies the expected utility axioms, it can be represented by the expected value of some function which is more or less unique.

The literature on expected utility theory is vast. It has been applied to all sorts of topics, and has received a great deal of defense, criticism, and mathematical elaboration.[5] Beyond a few remarks, this chapter will assume some sort of familiarity with the defense, but will rehearse many of the criticisms, particularly as they apply to ethics. We now need to ask: When is one lottery better than another? Which lotteries ought we to bring about? We begin with the first question.

[4] $L \succ M$ is defined as $L \succsim M$ and not $M \succsim L$. $L \sim M$ is defined as $L \succsim M$ and $M \succsim L$.

[5] At varying levels of philosophical and mathematical ambition, personal favourites include Fishburn (1970), Resnik (1987), Kreps (1988), Broome (1991), Hammond (1998), Ok (2007) and Gilboa (2009). In this volume, see Buchak (2016).

33.5 Expected goodness

Some philosophers imply that that if we know when one history is better than another, the question of when one lottery is better than another is straightforward. For example, Parfit (1984, p. 25) and Scheffler (1982, p. 1, note 2) start their discussions of consequentialism only by assuming

(1) The risk-free betterness relation is an ordering.

To cover risky cases, they think that we need to appeal only to expected utility theory. In particular, they think we just need to add

(2) One lottery is at least as good as another if and only if its expected goodness is at least as great.

In other words, the betterness relation is represented by the expected value of goodness. Parfit and Scheffler are not claiming that it is obvious when one history is better than another. Rather, they are claiming that once we have an ordering of histories in terms of betterness, (2) then tells us how to order lotteries in terms of betterness.

Now Parfit and Scheffler are quite brief about this and their real concerns lie elsewhere. But this sort of claim is commonly made, and it is important to realize that it contains a serious mistake. The basic difficulty is that (2) presupposes the existence of goodness measures, measures of how good histories are, and various problems arise depending on where we think these measures are coming from.

First, provided certain technical conditions are met, (1) guarantees that the risk-free betterness relation can be represented by some function.[6] To deal with the possibility that there may be more than one such function, we might treat the set of all goodness measures as the set of all of the functions which represent the risk-free betterness relation. It would then be natural to interpret (2) as saying: $L_1 \succsim L_2$ if and only if the expected goodness of L_1 is at least as great as the expected goodness of L_2 according to every goodness measure. Unfortunately, however, this approach leads to massive indeterminacy. An example will illustrate. Suppose there are exactly three histories x, y and z, ordered $x \succ y \succ z$ by the risk-free betterness relation. Let L be the lottery $[\frac{1}{2},x;\frac{1}{2}z]$ and let us consider how it compares with y. Consider the two functions u and v defined by $u(x) = v(x) = 1$, $u(y) = 0.9$, $v(y) = 0.1$, and $u(z) = v(z) = 0$. Both of these functions represent the risk-free betterness relation, and therefore count as goodness measures on the current proposal. But according to u, the expected goodness of L is less than that of y, and according to v, the expected goodness of L is greater than that of y. The current proposal therefore leaves L and y unranked, and it only takes a bit more work to show that this will be true of almost every pair of lotteries. So interpreting (2) along these lines does almost nothing to cover risky cases.

Secondly, to get around this problem we might hope to narrow down all of the functions which represent the risk-free betterness relation to (essentially) a single function to be used

[6] The result goes back to Cantor; for details, see any reasonably advanced book on utility theory, such as Kreps (1988) or Ok (2007).

as a goodness measure.[7] This line of thought is tacitly quite common, and what tends to happen is that one of the functions which represents the risk-free betterness relation seems quite simple or natural, and it is taken to be the goodness measure.[8] An old idea will illustrate. According to this idea, each "just noticeable difference" between outcomes is given the same magnitude of goodness, so that the difference in goodness between the best outcome and the second best outcome is equal to the difference in goodness between the second best outcome and the third best outcome, and so on.[9] In the toy example of the previous paragraph, this would be done by a function w where $w(x) = 1$, $w(y) = 0.5$, and $w(z) = 0$. Using (2) would then provide a ranking of all lotteries in terms of betterness. For example, L and y would turn out to be equally good. However, this proposal is ethically entirely arbitrary, and it is easy to invent circumstances in which the method delivers implausible conclusions. To illustrate, let us apply the same idea to individual betterness relations. Consider a wine connoisseur who is able to discriminate among a vast number of wines, and let us take her ordering of wines as given. Let a^+ be the outcome in which she gets the best possible wine, a the next wine down, r some rough house wine, and r^+ the next one up. The current method would regard the two lotteries $[\frac{1}{2}, a^+; \frac{1}{2}, r]$ and $[\frac{1}{2}, a; \frac{1}{2}, r^+]$ as equally good. But our connoisseur might regard experiencing the best possible wine as worth risking a lot for, and improving a rough house wine as hardly worth anything, leading her to conclude that the first lottery is better. But the current method woodenly regards the two lotteries as equally good.

Thirdly, one might approach the problem from a different direction. Suppose we start with a claim which is presupposed by (2), namely

Social EUT The betterness relation satisfies the expected utility axioms.

Now by the vNM theorem, Social EUT implies

(3) For some real-valued function on histories f, the betterness relation is represented by the expected value of f.

We might then define f as a goodness measure (along with its positive affine transformations). It follows that (2) now gives us the right results: one lottery is better than another just in case its expected goodness is greater. Unfortunately, however, just as the first method yielded almost complete indeterminacy, this method is almost completely uninformative. In almost all cases, it provides us with no concrete method of ranking lotteries. For example, in the toy example used to show why the first method leads to indeterminacy, it is consistent with the present method that L is better than y, that L and y are equally good, and that L is worse than y.

We have now looked at three ways of trying to fill in the story gestured towards by Parfit, Scheffler, and many others, the story which thinks that once we are given the risk-free

[7] More precisely, to a set of functions which are all related by positive affine transformation. The vNM theorem tells us that these will all be equivalent when it comes to ordering lotteries in terms of expected goodness.

[8] For example, McCarthy (2013) argues that this approach is common in accounts of the priority view and leads to unsatisfactory definitions of it.

[9] The basic idea goes back to Edgeworth (1881). For criticism and defense see e.g. Vickrey (1960) and Ng (1975) respectively.

betterness relation, we need only to appeal to expected utility theory to cover risky cases. Each attempt to say where goodness measures are coming from leads to a problem. The first leads to indeterminacy, the second to arbitrariness, and the third to uninformativeness. Now expected utility theory does indeed turn out to be a powerful tool for thinking about evaluative questions about risk, and even questions which do not seem to be about risk. But the story has to be more sophisticated than anything we have so far seen.

33.6 Veils of ignorance

To simplify, I will from now on assume that in evaluating lotteries, we are only concerned with the ethics of distribution, and in addition, not concerned with rights or responsibilities. In particular, I will assume: if h_1 and h_2 contain the same population and for each member i, h_1 is exactly as good for i as h_2, then h_1 and h_2 are equally good.

The best-known strategy for augmenting an appeal to expected utility theory is to use a so-called veil of ignorance, made famous but used in different ways by Harsanyi (1953) and Rawls (1971).

Assume a fixed population $1,\ldots,n$. Harsanyi's presentation of his argument tacitly identifies individual betterness relations with individual preference relations. But there are objections to that identification, and following Broome (1991) we can avoid them by restating Harsanyi's argument in terms of individual betterness relations. This enables us to leave it open whether the content of individual betterness relations has to do with preference satisfaction, the quality of experience, achievements, or some other account. Harsanyi's argument then begins with

Individual EUT Individual betterness relations satisfy the expected utility axioms.

Assume also that interpersonal comparisons are unproblematic in that

Interpersonal Completeness For all individuals i and j and histories h_1 and h_2, either h_1 is at least as good for i as h_2 is for j, or vice versa.

Together Individual EUT and Interpersonal Completeness imply that there are real-valued functions $u_1,\ldots,u_n$ on histories such that (i) for each individual i, i's individual betterness relation is represented by the expected value of u_i, and (ii) for all individuals i and j, h_1 is at least as good for i as h_2 is for j if and only if $u_i(h_1) \geq u_j(h_2)$. From now on, $u_1,\ldots,u_n$ will always be such functions, but their existence presupposes Individual EUT and Interpersonal Completeness. I will sometimes call them utility functions.

Harsanyi (1953) took ethics to be impartial.[10] But how should this be modeled, or made more concrete? This is where Harsanyi appeals to a veil of ignorance. Choosing under the

[10] Some of the arguments which follow make slightly stronger assumptions about interpersonal comparisons than I have made explicit. The point of these is to make various impartiality assumptions have an effect, and also to guarantee that the functions $u_1,\ldots,u_n$ are essentially unique, in that if some other set of functions $v_1,\ldots,v_n$ plays their role, there are real numbers $a > 0$ and b such that for all i, $v_i = au_i + b$. But I will suppress this slightly technical issue. For full details, see e.g. Broome (2004, p. 96).

equiprobability assumption is understood as choosing between two social situations on the assumption that one is equally likely to turn out to be each member of the population. Then Harsanyi took the idea that ethics is impartial to be well-modeled by

Veil of Harsanyi One lottery is at least as good as another if and only if it would be weakly preferred by every self-interested and rational person choosing under the equiprobability assumption.

I will skip the formal details, but from Individual EUT, Interpersonal Completeness and Veil of Harsanyi, Harsanyi gave a simple argument for

Sum The betterness relation is represented by the expected value of the function $u_1 + \cdots + u_n$.

Rawls (1971) agrees with Harsanyi that ethics is impartial, and that a veil of ignorance is a good way of modeling impartiality. To focus on their treatment of veils, we will ignore other differences, such as the different ways in which they understand interpersonal comparisons. With those aside, Rawls can be taken as agreeing with Individual EUT and Interpersonal Completeness. But his interpretation of the veil differs. Choosing under the *uncertainty assumption* is understood as choosing between two social situations on the assumption that one will turn out to be one of the members of the population, but with complete uncertainty about who that will be. Then Rawls took the idea that ethics is impartial to be well-modeled by

Veil of Rawls One history is at least as good as another if and only if it would be weakly preferred by every self-interested and rational person choosing under the uncertainty assumption.

Rawls then argued that Individual EUT, Interpersonal Completeness, and Veil of Rawls would result in

Maximin One history is better than another if and only if the former is better for the worst off.

Many commentators have thought Rawls should instead have concluded with

Leximin One history is better than another if and only if it is better for the worst off, or equally good for the worst off and better for the second worst off, and so on.

These arguments raise three basic questions: (i) What does rational choice under the uncertainty assumption really require? (ii) Given that one is going to model impartiality via some sort of veil of ignorance, is the uncertainty assumption a better way of doing it than the equiprobability assumption? (iii) Is modeling impartiality via a veil of ignorance a good idea anyway?

Briefly, (i) seems to be unclear. For example, suppose the Ramsey-Savage story is right about rational choice under conditions of uncertainty. For the agent behind the veil to lack implicit subjective probabilities of any degree of determinateness – and thus to model complete uncertainty – that story implies that her preferences are incomplete. At best, maximin (or leximin) would then seem to be but one rationally permissible

choice among many, whereas Rawls needs it to be rationally required (see Angner (2004) for further discussion). For (ii), the equiprobability assumption seems at first glance a reasonable attempt at giving impartiality a concrete and reasonably clear interpretation. Moreover, given the difficulties in understanding what rationality in conditions of complete uncertainty requires, it is hard to see what motivates shifting to the uncertainty assumption, aside from a question-begging attempt to avoid Sum. I will return to some of these issues, but the most fundamental question is (iii), and a later result of Harsanyi's seems to show that the use of veils of ignorance was never a good idea in the first place.

33.7 Harsanyi's theorem

To present Harsanyi's result we need to state two more premises. We continue to assume a fixed population. The first premise expresses a kind of impartiality.

Impartiality For all histories h_1 and h_2, if there is some permutation π of the population such that for each individual i, h_1 is exactly as good for i as h_2 is for $\pi(i)$, then h_1 and h_2 are equally good

The second premise is a so-called Pareto assumption.

Pareto (i) If two lotteries are equally good for each member of the population, they are equally good. (ii) If one lottery is at least as good for every member of the population and better for some members, then it is better.

This is Harsanyi's theorem. For an accessible proof, see e.g. Resnik (1987).

Theorem 2 *(Harsanyi) Assume a constant population. Then Individual EUT, Interpersonal Completeness, Social EUT, Impartiality, and Pareto jointly imply Sum.*

To recap what Sum says, the conclusion of the theorem says that one lottery is better than another just in case it has a greater sum of individual expected utilities. This implies that one history is better than another just in case it has a greater sum of individual utilities. However, in its classical form, utilitarianism is usually defined as the claim that one history is better than another just in case it has a greater sum of individual goodness. This raises the disputed question of what Sum has to do with utilitarianism, and thus whether Harsanyi's premises imply utilitarianism. Roughly speaking, Harsanyi's premises imply the classical version of utilitarianism just in case individual utilities are measures of individual goodness. Simplifying somewhat, Sen (1976) and Weymark (1991) denied that the two should be identified, whereas along with e.g. Harsanyi (1977b), Broome (1991), and Hammond (1991), I believe that they should be identified. I will say more about this in section 33.14, but the most important claim is that it does not really matter who is right. The conclusion of Harsanyi's theorem appears to tell us exactly what the content of the betterness relation is, and what name we should give to that conclusion is of much less importance.

In my view, it is hard to exaggerate the importance of Harsanyi's result. I will assume enough familiarity with expected utility theory, references to which were provided earlier,

to see the prima facie case for Individual EUT and Social EUT. The rough idea is that the prima facie case for rational preference relations satisfying the expected utility axioms can be modified to apply directly to evaluative relations like individual betterness relations and the betterness relation. The prima facie case for the other premises is fairly natural as well. The best way to explore this further will be to look at criticisms of the premises. We will do that shortly, but first I want to consider how Harsanyi's theorem improves on what we have seen so far.

The popular appeal to expected utility theory sketched in section 33.5 suffered from telling us little of any use about the betterness relation. But if we take individual betterness relations as given, and accept the premises of Harsanyi's theorem, the theorem shows that the content of the betterness relation is completely determined.

Consider now veil of ignorance arguments. Both Harsanyi's and Rawls's accept Individual EUT and Interpersonal Completeness. That leaves Harsanyi's veil argument with Veil of Harsanyi and Rawls's with Veil of Rawls, while Harsanyi's theorem is left with Social EUT, Impartiality, and Pareto.

Harsanyi's veil argument works by assuming that the person behind the veil is rational, and therefore has preferences which satisfy the expected utility axioms. Given that, Veil of Harsanyi yields Social EUT, and also, obviously, Impartiality and Pareto. So Harsanyi's veil argument enjoys no advantage over his theorem, and the theorem simply bypasses worries about veil arguments expressed by e.g. Scanlon (1982).

The comparison with Rawls is less clear. When discussing the veil, Rawls usually considers only the problem of ranking different histories. But someone behind the veil could also try to rank different lotteries (thus facing two forms of ignorance: uncertainty behind the veil, and risk beyond the veil). So we can ask what she thinks about Social EUT, Impartiality, and Pareto. It would be surprising if the uncertainty assumption led her to reject any of these claims, and thence Sum. But since Rawls is so plainly opposed to Sum, I think this suggests that aspects of his informal reasoning have not been fully captured in what seems to be his formal model. Sections 33.11 and 33.12 will discuss two major Rawlsian worries about some of Harsanyi's premises. But to foreshadow, these worries can be expressed directly as criticisms of the premises of Harsanyi's theorem, and appealing to the veil does not seem to add anything.

Finally, we will see in section 33.9 that there is at least one major view about the ethics of distribution which is impartial but is immediately ruled out by the adoption of a veil of ignorance, whether Harsanyi's or Rawls's. So much the worse for the veil as a model of impartiality. Thus in my view, the veil turns out to be just an unhelpful distraction, and the proper focus of attention for the ethics of distribution should be Harsanyi's theorem.

33.8 Variable populations

Before looking at various worries about and alternatives to the premises of Harsanyi's theorem, it is worth mentioning a way in which it can be extended. Problems where the population can vary are difficult. But we do not need to add much to the premises of Harsanyi's theorem to make progress.

The following says that only the kinds of lives people are living matters, not the identities of those people.

Anonymity For all histories h_1 and h_2 containing finite populations of the same size, if there is a mapping ρ from the population of h_1 onto the population of h_2 such that for every member i of the population of h_1, h_1 is exactly as good for i as h_2 is for $\rho(i)$, then h_1 and h_2 are equally good.

This premise makes the nonidentity problem discussed by Parfit (1984) rather trivial: if no one else will be affected, and a woman has to choose between having one of two different children, Anonymity plus Pareto implies that it would be better if she had the child whose life would be better.

Let U be the function defined on histories such that for any history h with population $1,\ldots,n$,

$$U(h) \coloneqq u_1(h) + u_2(h) + \cdots + u_n(h).$$

Then the premises of Harsanyi's theorem, but with Impartiality replaced by the stronger Anonymity, jointly imply

Same Number Claim Assume that all histories contain populations of the same size. Then the risk-free betterness relation is represented by U.

Turning to comparisons between populations of different sizes, I will outline an approach due to Broome (2004) and Blackorby, Bossert, and Donaldson (1995). I lack the space to discuss the details, but the crucial step is to argue for the

Neutral existence claim There exists a life l such that in every situation, provided no one already existing is affected, (i) it is better to create an extra life which is better than l; (ii) it is worse to create an extra life which is worse than l; (iii) it is a matter of indifference to create an extra life which is exactly as good as l.

Call such a life a *neutral existence*. Given a parameter v, let V be the function defined on histories such that for each history h with population $1,\ldots,n$

$$V(h) \coloneqq (u_1(h) - v) + (u_2(h) - v) + \cdots + (u_n(h) - v)$$

Some simple algebra shows that the same number and neutral existence claims together imply the

Variable number claim Assume that all histories contain finite populations. Then the risk-free betterness relation is represented by V, where v is the utility level of a neutral existence.

The value of v makes no difference to same number problems. For when comparing two histories with populations of the same size using V, the subtracted v's cancel out. In variable number problems, the presence of v in the definition of V means that ignoring effects on other people's lives, someone's existence makes a positive contribution towards goodness if and only if her life is better than a neutral existence.

Nothing so far said tells us what the value of v is, however. Setting it will involve further ethical issues, and is difficult to do in a way which respects common intuitions (Broome, 2004). For example, setting it low leads to the conclusion that a large number of people (e.g. a billion) all with extremely good lives is worse than an extremely large number (e.g. a billion billion) all with lives which may seem hardly worth living. Parfit (1984) evidently did not think much of this idea when he famously called it "the repugnant conclusion." On the other hand, setting the value of v high makes it bad to create someone who would have an intuitively good life, and that may seem implausible too.

When we ethicists first start to think seriously about probability, it may seem like a bane for us, vastly expanding the complexity of questions we have to address. But it may now look like a blessing. The problem of aggregating individual well-being to form an overall judgment about when one history is better than another seems difficult. Yet without appearing to make any assumptions about aggregation, and instead by largely appealing to expected utility theory, which is all about probability, Harsanyi's theorem seems to provide a solution. Section 33.15 will provide a closer look at the question of whether the theorem really does solve the "problem of aggregation." But we first examine criticisms of and alternatives to Harsanyi's premises which are also about probability.

33.9 Equality and fairness

The additive form of the conclusion of Harsanyi's theorem will make some suspect that its premises conflict with the idea that in the distribution of goods, equality and fairness matter. But where, if anywhere, is the tension? Assume a population of two people, A and B, and consider the following lotteries, which combine examples due to Diamond (1967) and Myerson (1981).

L_E	*heads*	*tails*
A	1	0
B	1	0

L_F	*heads*	*tails*
A	1	0
B	0	1

L_U	*heads*	*tails*
A	1	1
B	0	0

Anyone who thinks that equality is valuable should think that L_E is better than L_F. For while L_E and L_F are equally good for each person, L_E has in its favour that it guarantees equality of outcome while L_F guarantees inequality (Myerson, 1981). But Pareto implies that L_E and L_F are equally good, so it is inconsistent with the idea that equality is valuable.

Anyone who thinks that fairness is valuable should think that L_F is better than L_U. For while Impartiality implies that the outcomes under L_F and L_U are equally good, L_F has in its favour that it distributes the chances fairly (Diamond, 1967).

Diamond's example leads to the first of a series of challenges to the assumptions about expected utility in Harsanyi's premises. By Impartiality, all of the outcomes under L_F and L_U are equally good. Strong Independence of the betterness relation then implies that L_F and L_U are equally good.[11] Hence the assumption that the betterness relation satisfies the expected

[11] Proof: for all lotteries L and M, write $L \gtrsim M$ for "L is at least as good as M". By Impartiality, $[1,0] \sim [0,1]$. Strong Independence for $\gtrsim$ then implies $L_U = \frac{1}{2}[1,0] + \frac{1}{2}[1,0] \sim \frac{1}{2}[0,1] + \frac{1}{2}[1,0] = L_F$ as required.

utility axioms, in particular Strong Independence, clashes with the idea that fairness is valuable.

I think that Myerson's and Diamond's examples lie at the heart of concerns with equality and fairness.[12] It is difficult to argue for this in a short space, though section 33.16 will say more. But suppose it is correct. How could the examples be generalized into full-blown theories about what it is for equality or fairness to be valuable?

I will just illustrate an approach for the case of equality. Suppose we are given a preorder $\succsim_e$ on histories such that $h_1 \succsim_e h_2$ if and only if h_1 is uncontroversially (among egalitarians) at least as good in terms of equality as h_2. My own account of the extension of $\succsim_e$ is in McCarthy (2015). But to give two simple cases, every equal distribution is going to be uncontroversially better in terms of equality than every unequal distribution, and all equal distributions are going to be uncontroversially equally good in terms of equality. Consider

Equality-neutral Pareto Assume a fixed population. For all lotteries $L_1 = [p_1, h_1; \ldots p_m, h_m]$ and $L_2 = [p_1, k_1; \ldots; p_m, k_m]$: (i) *if* L_1 is exactly as good as L_2 for all individuals, and $h_j \sim_e k_j$ for all j, *then* L_1 and L_2 are equally good; and (ii) *if* L_1 is at least as good as L_2 for all individuals and better for some individual, and $h_j \succsim_e k_j$ for all j, *then* L_1 is better than L_2.

Equality principle Assume a fixed population. For all lotteries $L_1 = [p_1, h_1; \ldots; p_m, h_m]$ and $L_2 = [p_1, k_1; \ldots; p_m, k_m]$: *if* L_1 is at least as good as L_2 for all individuals, $h_j \succsim_e k_j$ for all j and $h_j \succ_e k_j$ for some j, *then* L_1 is better than L_2.

McCarthy (2015) argues that together, these principles are the core of egalitarianism. Equality-neutral Pareto is a weakening of Pareto, designed to avoid clashes with examples like Myerson's. The equality principle is designed to generalize the idea that equality is valuable, as illustrated by Myerson's example. Thus we obtain a very general egalitarian theory by starting with Harsanyi's premises, weakening Pareto to its equality-neutral cousin, then adding the equality principle.

Notice that the equality principle is inconsistent with the adoption of either Harsanyi's or Rawls's veil of ignorance. But it can easily be shown to be consistent with the notion of impartiality captured by Impartiality. So if it was meant only to model impartiality, the adoption of a veil of ignorance is too strong.

The characterization of the idea that equality is valuable via the equality principle exploits natural dominance ideas. Roughly speaking, suppose that each part of some object x is at least as good with respect to some value V as the corresponding part of object y. Then x is said to *weakly dominate* y in terms of the value V. If x weakly dominates y, but y does not weakly dominate x, then x *strictly dominates* y. Thus the equality principle says that if L_1 weakly dominates L_2 in terms of well-being, and strictly dominates L_2 in terms of equality, then L_1 is better than L_2. I lack the space to discuss the details, but I believe that the way to characterize the idea that fairness is valuable is to develop dominance ideas in a way suggested by Diamond's example. However, while the apparent similarities between Diamond's and Myerson's examples suggest parallels, it appears that there are subtle asymmetries between concerns with equality and concerns with fairness (McCarthy and Thomas, 2016).

[12] This is not quite right. In my view it is better to say that Myerson's example is about equality of outcome, and Diamond's is about equality of prospects, not fairness. But here I stick with the more usual terminology. For reasons for not talking about fairness, see McCarthy (2015).

33.10 Priority

Parfit (2000) argued that what he called the priority view is an important alternative to egalitarianism, sharing many of its apparent virtues but avoiding what he called the leveling-down objection. He summarized it via the slogan that "benefiting the worse off matters more", but commentators have been divided over whether he managed to articulate a genuine alternative to egalitarianism.

A puzzle about making sense of the priority view is that its distinctive feature is advertised as an intrapersonal phenomenon: what is bad about people being worse off is that they are worse off than they might have been (Parfit, 2000, p. 104). This has suggested to commentators that according to the priority view, it matters more to more to benefit someone the worse off she is even when no others are around at all (Rabinowicz, 2002). But in cases where only one person is around and risk is not involved, the priority view, like any other sane view, will accept that one history is better than another if and only if it is better for the sole person.

Matters are different, however, when risk is involved. Several commentators have thought that the priority view should be formulated in a way which makes it have distinctive consequences in one-person cases involving risk (Rabinowicz, 2002; McCarthy, 2008; Otsuka and Voorhoeve, 2009). I am inclined to go further and say that the key idea behind the priority view receives its clearest and most fundamental expression in such cases.

To illustrate, suppose A is the only person around, and compare the history $h = [1]$ with the lottery $L = [\frac{1}{2}, 2; \frac{1}{2}, 0]$, with the numbers supplied by u_A. Because L and h are equally good for A, Pareto implies that they are equally good. But I believe that the priority view should be understood as saying that h is better than L.

More generally, I believe that the key idea of the priority view is what I call the

Priority principle Assume a fixed population. Suppose histories h_1, h_2 and h_3 each contain perfect equality. Then (i) h_1 is at least as good as h_2 if and only if h_1 is at least as good for each individual as h_2; and (ii) *if* for each individual i, h_1 is better for i than h_2, h_2 is better for i than h_3, and h_2 is exactly as good for i as $L = [\frac{1}{2}, h_1; \frac{1}{2}, h_3]$, *then* h_2 is better than L.

Notice that this is inconsistent with equality-neutral Pareto. Some writers find it absurd that in one-person worlds, the betterness relation and the sole person's individual betterness relation could diverge (e.g. Otsuka and Voorhoeve, 2009), as the priority principle implies. Rabinowicz (2002) regards this claim as acceptable, while Parfit (2012), for example, offers a defense.

But rather than discuss possible defenses of the priority principle, I will note a less discussed objection to the priority view. The priority view can be formulated by starting with Harsanyi's premises, weakening Pareto far enough to accommodate the priority principle, then adding the priority principle (McCarthy, forthcoming a). But when this is done, any account of the extension of the betterness relation which is consistent with the Harsanyi premises turns out to be consistent with the priority view premises, and vice versa. But the priority view has a more complicated way of describing the betterness relation, because of the less simple relationship it posits between betterness and individual betterness in one-person worlds. So the objection is that the priority view fails to provide a reasonable alternative to the Harsanyi premises, not because of any ethically absurd implications, but

because of the theoretical vice of needless complexity (cf. Harsanyi, 1977b; Broome, 1991; McCarthy, 2008, forthcoming a).[13]

33.11 Continuity

Continuity is seldom discussed. When it is mentioned, it is often said just to be a technical assumption. But when the claim is that the betterness relation or individual betterness relations satisfy Continuity, this is a clear mistake.

To illustrate, let a be a very good life, a^+ a slightly better life, and z an extremely bad life, such as being in severe pain or enslaved for a long time. The claim that individual betterness relations satisfy Continuity implies that there is a gamble which would almost guarantee an individual a^+ with a small chance of z which is better for the individual than having a for certain. But regardless of what one thinks about this case, it is not a technical assumption to claim that the risk is worth it. It is a substantive evaluative judgment, and different views about it are reasonable. For what it is worth, I believe that many of Rawls's informal remarks about his veil of ignorance would have been more naturally modeled by denying that individual betterness relations satisfy Continuity because of this kind of case than by his actual model.

It is clear that Continuity is something ethicists should pay attention to. The good news is that the result of weakening the expected utility axioms by dropping Continuity is formally well understood, thanks to results by Hausner (1954) and others.

But there are several pieces of bad news. First, the general statement of Hausner's result is quite mathematically complex and not easy to speak about informally. Secondly, it is time to stop speaking of *the* continuity axiom. There are several EUT-style continuity axioms (see e.g. Hammond, 1998), and it is far from clear what the ethical grounds for adopting one but not another might be. Thirdly, speaking loosely, Continuity failures occur when one lottery in some sense has "infinitesimal" value compared with another. But such cases pose a challenge to standard treatments of probability as well, and this needs to be incorporated into the analysis.[14] In summary, perhaps in the end ethicists can safely ignore Continuity. But it would be better to know that than to hope for it, and the work needed to arrive at such a conclusion appears to be substantial.[15]

33.12 Incommensurability

One of the major contributions of the contractualist literature has been to force us to take seriously difficulties with evaluative comparisons of different kinds of goods. But part of

[13] As an analogy, consider again the best-system analysis of laws. Suppose someone offers some account of the laws of the world which captures all relevant facts. But this account is more complex than some other account which also captures all relevant facts. On the best-system analysis, the more complex account is mistaken about what the laws are, despite getting the relevant facts right. McCarthy (forthcoming a) argues that the priority view is mistaken on similar grounds.

[14] For an accessible account of how the challenge applies to Savage's treatment of subjective probability, and a sketch of mathematically sophisticated responses, see Gilboa (2009) pp. 99–100.

[15] For recent work in this direction, see Jensen (2012).

the assumption that the betterness relation and individual betterness relations satisfy the expected utility axioms is that these relations are complete. But from the perspective of difficulties with evaluative comparisons, such completeness assumptions look far from obvious. They may seem particularly implausible if we adopt the popular view that the basis for such things as interpersonal comparisons should be as neutral as possible between competing substantive views about what a good life is, as argued, for example, in Rawls (1982).

One response would be to adopt something like resources, freedoms, or opportunities as the basis for interpersonal and intrapersonal comparisons (see e.g. Rawls, 1982; Sen, 1985). However, the premises of Harsanyi's theorem are silent on the content of individual betterness relations, so there is no obvious reason why the theorem cannot be run when their content is understood in terms of resources and so on. Nevertheless, even resources have their own problems to do with comparability because of the different nature of different kinds of resources. So this response is a diversion, and we should turn directly to Harsanyi's premises to see what can be done about difficulties with comparability.

The most immediately tempting response is simply to drop the completeness assumptions. This means that the various evaluative relations featuring in the theorem become preorders which are not assumed to be complete. A large advantage of working with preorders is that mathematically speaking, they are relatively tractable. For example, a corollary of Szpilrajn's theorem is that a preorder is identical to the intersection of all of the complete preorders which extend it. This has the advantage that in thinking about preorders one can often work with complete preorders anyway.

This corollary is strikingly parallel to the supervaluationist treatment of vague predicates: a sentence involving a vague predicate is true if it is true on all admissible sharpenings of the predicate, false if it is false on all admissible sharpenings, and neither true nor false otherwise.

But this should suggest caution: if a natural response to difficulties to do with comparability is to shift to preorders, the response looks like one of the classic candidates for a solution to the problem of vagueness. But supervaluationist approaches have been heavily criticized (see e.g. Williamson, 1994). Furthermore, perhaps the parallel suggests that the basic problem with comparing different kinds of goods is one of vagueness. In fact, cases in which evaluative comparisons look extremely difficult seem to lend themselves to sorites paradoxes, one of the hallmarks of vagueness.

In one way this is good news: there is a vast amount of work on vagueness, so ethicists have plenty of material to borrow from. Since the topic is probability, it is worth mentioning that some treatments of vagueness are probabilistic, and that an extensive literature takes this approach to vague comparatives; see e.g. Fishburn (1998) for a survey. In another way it's bad news: perhaps the main reason why there is so much literature on vagueness is the almost complete lack of consensus.

Perhaps we ethicists should just shelve the problem of how best to model difficulties to do with evaluative comparisons until there is more convergence in the literature on vagueness. However, in the absence of such convergence, it may still be possible to achieve some kind of stability result: show that the solutions to a class of interesting ethical problems which involve goods which are difficult to compare are insensitive to the resolution of more general problems about vagueness. For example, Broome (2004) takes this approach in his discussion of the neutral level for existence. In section 33.16 I will suggest that the same can be done for the question of what Harsanyi's theorem really shows.

33.13 Non-expected utility theory

The backbone of Harsanyi's theorem is expected utility theory, but we have seen a number of ways in which the claim that various evaluative relations satisfy the expected utility axioms can be criticized. The axioms so far criticized are Strong Independence, Ordering (insofar as completeness was criticized), and Continuity. Some writers even go so far as to criticize transitivity (see e.g. Temkin, 2012).

These criticisms are directly based either on distributive intuitions (Strong Independence, Continuity), or on the nature of goods being distributed (Ordering). But a serious question about the expected utility axioms arises from a different direction.

Since Allais (1953) and Ellsberg (1961), it has appeared to many that individual preference relations violate the expected utility axioms in fairly systematic ways. The attempt to describe these violations has led to a huge body of work developing alternatives to the expected utility axioms (for surveys see e.g. Schmidt, 2004; Sugden, 2004; Gilboa, 2009; Wakker, 2010).

This project has been accompanied by two broad views. One is that the alternative axioms simply help us catalogue human irrationality, which might of course be very important in various descriptive and explanatory contexts. The other, often prompted by the fact that the violations are often stable under criticism, is that the support the alternative axioms tacitly enjoy genuinely threatens the picture of rationality provided by expected utility theory.

Now these are views about rationality, whereas we have been interested in such things as betterness and betterness for people. But the development of non-expected utility theory suggests that it would be interesting to modify distributive theories which to varying extents involve the expected utility axioms by weakening those axioms and then adding some of the non-expected utility axioms.

If the application of the non-expected utility axioms to such things as individual betterness relations turns out to be reasonably well motivated, the result should be an expanded account of reasonable distributive theories. But even if those axioms are not well motivated when applied to evaluative relations, this project would still be worth pursuing. If a class of popular distributive intuitions turns out to be generated by such an application of non-expected utility theory, we would in effect have an important error theory.

33.14 Evaluative measures

Discussions of the ethics of distribution commonly assume the existence of quantitative measures of various evaluative properties, then use these measures to formulate various apparently natural ideas. For example, individual goodness measures, quantitative measures of how good histories are for individuals, are often taken to exist. Then assuming a constant population $1, \ldots, n$, it is often claimed that

(**U**) According to utilitarianism, two histories are equally good if they contain the same sum of individual goodness.

(**E**) According to egalitarianism, an equal distribution is better than an unequal distribution of the same sum of individual goodness.

(**P**) According to the priority view, it is better to give a unit of individual goodness to a worse-off person than to a better-off person.

These claims tacitly assume that talk of units of individual goodness is well-defined. They are often taken to be (at least partial) definitions of the distributive theories in question, making what seems natural or appealing about the theories in question transparent. For more detail, McCarthy (2013) examines the role of evaluative measurement in common understandings of the priority view.

However, there are serious difficulties with this kind of approach to the ethics of distribution. I will mention just one specific problem.

The only obvious fact about individual goodness measures is that they have to represent risk-free individual betterness relations. But this only makes individual goodness measures unique up to increasing transformation.[16] But for units of individual goodness to be well-defined, individual goodness measures must be unique up to positive affine transformation. So to make them well-defined it looks as if we need to make an arbitrary choice of measure (Broome, 2004, p. 90). But this will make the theories partially defined by (U), (E), and (P) rest on an arbitrary choice, and fail to vindicate the idea that they are the fundamental theories about the ethics of distribution we take them to be.

More generally, taking the existence of quantitative evaluative measures as given, then using them to theorize about the ethics of distribution, is strongly at variance with standard views about measurement in the physical and social sciences. There, quantitative measures are seen as emerging as canonical descriptions of qualitatively described prior structures (see e.g. Krantz, Luce, Suppes and Tversky, 1971; Narens, 2007; Roberts, 2009). My own view is that we should treat evaluative measurement along the same lines.

By itself, this does not begin to settle what we should say about individual goodness measures. But individual goodness measures turn out to be well-enough defined for talk of units, sums, and so onto make sense, at least given certain background assumptions.

I can only sketch this view, but in more detail, sections 33.9 and 33.10 point to a characterization of egalitarianism and the priority view in terms of primitive qualitative relations (betterness, individual betterness). Similarly, I think the premises of Harsanyi's theorem should be understood as characterizing utilitarianism. Now (U), (E), and (P) are close to platitudinous. But given these characterizations of utilitarianism, egalitarianism and prioritarianism, this means that we can treat (U), (E), and (P) as implicit definitions of individual goodness measures. The result is that individual goodness measures turn out to be the positive affine transformations of $u_1, \ldots, u_n$, or what Broome (1991) calls Bernoulli's hypothesis. For details, see, for example, McCarthy (2015).

The background assumptions are that individual betterness relations satisfy the expected utility axioms and that interpersonal comparisons are unproblematic. But what if these fail? I will not pursue this, for I think the most important lesson about evaluative measures is not that they are arguably well-defined, but that it does not much matter. We can and

[16] I.e. if some function f represents the risk-free betterness relation, and g is some strictly increasing function on the reals ($x < y \implies g(x) < g(y)$), then $g \circ f$ also represents the risk-free betterness relation.

should theorize about the questions which really matter in the ethics of distribution without using evaluative measures. By focusing instead on comparatives and various claims about probability, none of the distributive views we have been discussing presuppose the existence of evaluative measures, the preeminent example, of course, being Harsanyi's.

33.15 Aggregation

But this raises the question of what Harsanyi's theorem really shows. Ethicists often talk about the "problem of aggregation". What they typically have in mind is the task of somehow combining an assessment of what things are like for each individual in a particular situation to form some sort of overall judgment of the situation which enables us to make an evaluative comparison with other situations.

Supposing the premises of Harsanyi's theorem are correct, it is tempting to think that Harsanyi's theorem solves the problem of aggregation. I believe this was Harsanyi's view, and I think it is popular among welfare economists. Harsanyi did not use the terms 'individual betterness relation' and 'betterness relation', and I stress that the following passage is mine, not his. But I think the following captures the spirit of his view (see especially Harsanyi, 1977a).

> Determining the content of individual preferences relations (despite filtering out various irrationalities, excluding such things as sadistic preferences, and requiring preferences to be rich enough to enable interpersonal comparisons) is basically a psychological matter (Harsanyi, 1977a). It does not involve any significant evaluative or aggregative assumptions. But we should identify individual betterness relations with individual preference relations. Given the truth of Harsanyi's premises, Harsanyi's theorem then explicitly determines the extension of the betterness relation. Problem of aggregation solved.

This position underplays the role of evaluative assumptions in determining the content of individual betterness relations in at least two ways. First, determining the content of individual preference relations may well involve prior evaluative assumptions because of the role of such assumptions in popular accounts of radical interpretation (see e.g. Lewis, 1974). Secondly, even when they are restricted to histories, identifying individual betterness relations with individual preferences relations is highly controversial. It is a major evaluative question whether to understand the content of risk-free individual betterness relations in terms of preferences, the quality of the individual's experiences, her achievements, or some combination thereof.

But suppose that evaluative question has been settled, and that Harsanyi's premises are true. The theorem certainly shows that figuring out the content of the betterness relation is no harder than determining the content of individual betterness relations. But what exactly does it show about the problem of aggregation?

First, it is a vast exaggeration to say that the theorem solves the problem of aggregation. Problems of aggregation arise whenever we have to make some sort of assessment of a whole based on an assessment of its parts. But figuring out the content of individual betterness relations involves major questions of aggregation. Even in the case in which all outcomes are equally likely, to assess whether facing some lottery is better for someone than some particular outcome, we will have to assess what each of the possible outcomes of the lottery

are like for her, then somehow aggregate to reach an overall assessment of the lottery. This problem is complicated and is, in my view, much neglected. Like many economists, Harsanyi's own account tacitly appeals to the individual's preferences. But this should not seem very appealing to those of us who think that preference satisfaction accounts are mistaken even for the question of when one outcome is better for an individual than another.

Secondly, there is no logical reason why we cannot use the theorem to deduce the content of individual betterness relations from the content of the betterness relation, in particular from judgments about when one history is better than another. In cases where we are very confident about the latter, this will even seem appealing. I am afraid I lack the space to discuss this, but I think this idea provides a natural way of interpreting various contractualist comments about veil of ignorance arguments (see e.g. Scanlon, 1982; Nagel, 1970), in particular leading to an interesting case for rejecting the claim that individual betterness relations satisfy Continuity.

More generally, if its premises are true, Harsanyi's theorem teaches us that determining the content of the betterness relation is easier than we may have thought. But the flipside is that determining the content of individual betterness relations is harder than many of us have assumed.

33.16 Summary on evaluation

When thinking about the ethics of distribution, it may seem that the real evaluative questions are about when one history is better than another, or better for some individual. Factoring in probability may then seem like a basically technical exercise, not one ethicists need be much concerned with.

Almost every topic discussed could easily have its own survey article. I have had to omit many important positions, and give only sketchy defenses of positive positions. Nevertheless, I have tried to make the case for the opposite view. Not only are there very important ethical issues about how to rank lotteries, but these issues directly bear on questions about when one history is better than another. I will end the evaluative discussion with two opinions.

First, if I am right, almost every major position on the ethics of distribution is essentially to do with probability. For example, assuming a constant population $1 \ldots n$, concerns with fairness, equality, and giving priority to the worse-off as characterized in sections 33.9 and 33.10 can each be shown to be consistent with the popular idea that the risk-free betterness relation is represented by $w \circ u_1 + \cdots + w \circ u_n$ for some strictly increasing and strictly concave function w. These views come apart only when probability is introduced. So one aspect of the importance of probability is the increase in expressive power its introduction provides: it allows us to draw distinctions which are difficult or impossible to draw in a risk-free framework.

Secondly, I think the various challenges to Harsanyi's premises stemming from appeals to equality, fairness, priority, and non-expected utility theory fail. To be sure, there is at least a reasonable case for rejecting Continuity, and Ordering (at least, the completeness part of it) is under serious threat. Nevertheless, we can drop Continuity and under many ways of modeling difficulties to do with comparability, what I take to be the core lesson of Harsanyi's

theorem remains stable:[17] determining the content of individual betterness relations and determining the content of the betterness relation are just different descriptions of the same problem. This may help. Our initial judgments about individual betterness and about betterness may be in tension with each other, and we may be more confident about some judgments than others. Harmonizing these judgments in an attempt to achieve reflective equilibrium may increase our confidence in the result.

33.17 Ought

Expected utility theory has turned out to be hugely important for developing a taxonomy of answers to the fundamental evaluative question: when is one history or lottery better than another? I have not emphasized this, but I also think that the clarity of this taxonomy is also extremely helpful for assessing which answer is correct. In the remaining space I have room for only one suggestion which, though hardly very original, is that the same turns out to be true for decision theory and the fundamental normative question: what ought we to do?

One immediate disclaimer is needed. Expected utility theory is usually understood as a theory about the structure of the preferences of ideally rational agents. But this chapter has discussed the application of expected utility theory to understanding evaluative comparatives without having to say anything about rationality. Rather, many of the ideas and criticisms of expected utility theory are directly applicable to questions about evaluative comparatives.

Similarly, decision theory is usually understood as an account of ideally rational action, and it is typically assumed that the rationality of an action depends in some way upon the agent's preferences. However, we can apply many ideas from decision theory directly to questions about the fundamental normative question without having to presuppose some grand connection between rationality and ethics. For example, it is a serious mistake to think that decision theory is going to be important to ethics only if ethics is somehow about preference satisfaction, or if we hitch ourselves to the unlikely project of deriving ethics from rationality. Thus the discussion of decision theory in what follows is only meant to draw parallels between questions about ethics and questions about rationality. Because the debates about rationality are often better developed, these parallels may be illuminating. With no attempt at exhaustiveness, the sequel will look briefly at three examples, with particular emphasis on probability.

33.18 Act consequentialism

Given some account of betterness, the most obvious ethical theory is act consequentialism: what we ought to do is to bring about the best available lottery. If we assume for simplicity that the betterness relation satisfies the expected utility axioms, act consequentialism then

[17] In fact, this is true even if we weaken some of the EUT ideas in Harsanyi's framework and add various well-known nonEUT ideas. This is further pursued in McCarthy, Mikkola, and Thomas (2016) and McCarthy (forthcoming b).

implies that there is some value function such that we ought to perform the action with the greatest expected value. Thus act consequentialism is the ethical theory which most obviously parallels decision theory.

Act consequentialism is also one of the most criticized theories, one standard criticism being that it has implausible implications. For example, assuming an impartial method of valuation, Williams (1973) argued that act consequentialism undermines the partiality which for many people makes life worth living: devotion to personal projects and particular people, often friends and family. But this raises the question of what act consequentialism really requires in the first place.

Taking for granted a probability-based view which uses subjective probabilities, or at least, probabilities which are relative to the evidence available to the agent, Jackson (1991) famously argued that because of facts about each individual's probabilities, act consequentialism will typically not require each agent to promote general well-being and pursue whichever projects are the most impartially valuable. Rather, it will require a typical agent – Alice, let's call her – to promote the well-being of the relatively small group of people Alice knows and cares about, and to adopt and then pursue projects in which Alice takes a natural interest. This does not amount to a rejection of impartial valuation, but instead reflects facts about each agent's limited information, the costs of deliberation and of acquiring new information, the complexity of the interpersonal and intrapersonal coordination problems she faces, the effects her actions will have on the expectations others will have of her future behaviour, her motivational strengths, and so on. Such facts will be encoded in the agent's probabilities, and will therefore affect which of her acts will maximize expected value. Very often, Jackson argued, such acts will favour her nearest and dearest.

Jackson's argument was offered as a response to Williams, but it offers a much more general lesson. Understanding what act consequentialism implies is going to require sophisticated thinking about probability. The huge complexity of this problem stands in sharp contrast to the occasional complaint that act consequentialism is simple-minded.

33.19 Rule consequentialism

Many writers, however, prefer rule consequentialism (or contractualism: at the normative level, these views are often very similar). On the one hand, rule consequentialism seems to fit better with common opinion about what we ought to do than act consequentialism (it is said to secure rights etc.). On the other, it seems to avoid the obscurities of deontology by resting its account of what we ought to do on an appeal to what is good for people. But how is this achieved?

Harsanyi's writings on rule utilitarianism offer a relatively clear answer. Simplifying slightly, Harsanyi (1980) claims that each member of a society of act utilitarians will always maximize the sum of expected individual utilities where the calculation is based on her subjective probabilities of what the other members are going to do. Each member of a society of rule utilitarians is committed to and thus will always act upon the rule *R* which is such that if everyone acts according to *R* expected utility will be greater than if everyone acts according to some other rule (I ignore the possibility that two rules could be tied).

Harsanyi claims that rule utilitarianism will lead to "incomparably superior" overall results in comparison with act utilitarianism because of its superiority in two kinds of scenarios: (i) in certain simultaneous coordination games (e.g. choosing whether to vote), and (ii) in certain sequential games (typically involving choices about respecting rights, keeping promises etc.). This superiority is despite the fact that *R* will sometimes tell agents to perform actions which they are certain will produce suboptimal results, where optimality is understood in terms of maximizing the sum of expected utilities. This last feature leads many to suspect that there is something unstable about rule utilitarianism, but Harsanyi claims that these superior overall results imply that rule utilitarianism is correct.

It would take a separate article even to outline the important issues here, and I merely want to make three points to illustrate the potential value of looking at this style of argument through the lens of contemporary debates about decision theory. To do that, I will assume for the sake of argument (though this is far from obvious) that Harsanyi is right about the superior overall results of rule utilitarianism in comparison with act utilitarianism.

First, Harsanyi stresses that the rule utilitarians take themselves to be facing a problem involving complete probabilistic dependence: each will commit to (and thus act on) rule *R* if and only if all commit to *R*. In this respect, rule utilitarians are like clones in the well-known case of clones playing a prisoner's dilemma. It is this probabilistic dependence which leads to rule utilitarianism's superior performance in the coordination games. However, in these coordination games, there is causal independence between the actions of each player. But "probabilistic dependence yet causal independence" takes us to a crucial issue in decision theory. Very roughly, so-called *evidential decision theory* assesses (the rationality of) actions in terms of how likely good outcomes are conditional upon the actions being performed. By contrast, *causal decision theory* assesses actions in terms of their causal tendency to produce good outcomes. The classic case in which the two come apart is Newcomb's problem. However, for those of us who think that Newcomb's problem teaches us to be causal decision theorists (see e.g. Joyce, 1999), probabilistic dependence is a red herring when there is causal independence, as there plainly is in Harsanyi's simultaneous coordination games. So we may think that Harsanyi has tacitly built something like evidential decision theory into rule utilitarianism, and so much the worse for rule utilitarianism.

Secondly, the success of rule utilitarianism in various sequential games stems from the rule utilitarians' commitment to the rule *R* even in contexts in which acting on *R* leads to suboptimal results. The conclusion that in virtue of this success, rule utilitarianism is right about what we ought to do is parallel to a revision to standard decision theory later urged by Gauthier (1994) and McClennan (1990). This revision claims that if it is rational at time t to become committed to performing some action at a later time t' which is obviously irrational when considered in isolation, it is rational to commit to the action and then later perform that action. But those of us who take the toxin puzzle of Kavka (1983) to dramatize why this revision is mistaken may think that rule utilitarianism is making the same kind of mistake.

Thirdly, Harsanyi's characterization of act versus rule utilitarianism parallels the influential distinction in von Neumann and Morgenstern (1944) between games against nature and games against other people. Each act utilitarian will have probabilities about a number of relevant variables, and will maximize expected value accordingly. The fact that some of these variables are the behaviour of other people who like herself are act utilitarians is neither here nor there; the decision theoretic model still applies. But when an agent is in a situation in which the outcome depends in part on the behaviour of agents just

like her, von Neumann and Morgenstern argued that decision theory is inappropriate. The problem of self-reference embedded into such situations requires the different tools of game theory, and Harsanyi's rule utilitarians reason along similar lines. Perhaps von Neumann and Morgenstern's argument could be used to bolster Harsanyi's approach. Alternatively, those of us who are convinced by Skyrms (1990) in thinking that problems of self-reference can and should be handled without having to abandon decision theory may think this points to a further difficulty for rule utilitarianism.

Of course, the fact that Harsanyi focussed on rule utilitarianism rather than rule consequentialism has been inessential to the discussion. These crude and preliminary remarks are meant only to suggest the value of looking at the foundations of rule consequentialism through the lens of parallel and often much more extensive debates about decision theory.

33.20 Deontology

Those with strong deontological intuitions may reject rule consequentialism, either because they are not convinced that it is a stable alternative to act consequentialism, or because its conclusions are not deontological enough. But we may now seem to have reached the limits of the usefulness of thinking about decision theory. Very roughly, anything like a decision theoretic approach to deontology looks like the wrong model: the former is all about weighing goods against evils, and the latter thinks there are circumstances in which such weighing is illegitimate, or counts for nothing. Nevertheless, one lesson from thinking about probability is that weighing is not so easy to avoid.

In trying to characterize a deontological view, there seem to be two basic options. What I will call *agent-centered* views typically prohibit actions which would involve the agent's mental states bearing some kind of inappropriate relation to the outcome. The most obvious example is the so-called *principle of double effect*, which in its simplest form prohibits bringing about intended harm, but permits certain otherwise identical cases of bringing about merely foreseen harm. What I will call *causal structure* views typically prohibit actions which stand in some kind of inappropriate causal relation to the outcome. For example, in the famous trolley problem, an out-of-control trolley is going to kill five people who are stuck on the track, but a bystander can switch the trolley to a sidetrack where it will kill one person. Many people who have strong deontological intuitions think it is permissible to switch the trolley. But in most cases, they think that killing one to save five is impermissible, as in the variant where the bystander can push a fat man off a bridge to stop the trolley (Thomson, 1976), killing him but saving the five. Causal structure theorists think the intentions of the bystander are irrelevant, and search for differences in the causal structure of the cases to explain the difference in permissibility.

Many deontologists have not had much sympathy for agent-centered views, and have preferred some kind of causal structure view (e.g. Kamm, 1996). But here is what I believe is a relatively neglected problem about views. If the inappropriate causal relation is between the action and the outcome – as in, e.g., the fat man variant but not the trolley problem itself – then prima facie, there are going to be actions which bring about the following lotteries:

some benefit occurs with nonzero probability p, some inappropriate causal structure obtains with probability $1 - p$. For example, driving a truck across the bridge will either miss the fat man and deliver aid elsewhere, or else hit him and topple him off the bridge, stopping the trolley and saving the five.

What should causal structure deontologists say about such actions? There are at least five responses: (i) All such actions are impermissible. Objection: this leads to an intolerably restrictive view. (ii) Such actions are impermissible if and only if they turn out to result in the inappropriate causal structure. Objection: similar to the objections to outcome-based views in section 33.2. (iii) Actions which lead to the inappropriate causal structure with probability one are impermissible, all others are permissible. Objection: it is not credible that there should be such a gulf between probability one and probabilities just less than one. (iv) Actions performed by agents whose reasons for performing them include the benefits resulting from the inappropriate structure are impermissible. Objection: this collapses causal structure views into agent-centered views. (v) Actions are impermissible if and only if p exceeds some intermediate probability threshold. Objection: this seems to be the most principled response for a causal structure view, but it suggests the acceptability of weighing the alleged badness of the causal structure against the production of benefits. This seems to fit poorly with the guiding deontological image of the inappropriateness of weighing when inappropriate causal structures are concerned.

Perhaps this kind of case points towards a serious problem for causal structure views; see further Jackson and Smith (2006). Or it may provide an opportunity for causal structure theorists to refine their views. Either way, thinking about probability and deontology seems helpful.

Acknowledgments

Thanks to Alan Hájek and Kalle Mikkola for very helpful comments. Support was partially provided by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (HKU 750012H).

References

Allais, M. (1953) Le comportement de l'homme rationnel devant le risque, critique des postulates et axiomes de l'ecole américaine. *Econometrica*. 21. pp. 503–46.

Angner, E. (2004) Revisiting Rawls: *A Theory of Justice* in the light of Levi's theory of decision. *Theoria*. 70. pp. 3–21.

Blackorby, C., Bossert, W., and Donaldson, D. (1995) Intertemporal population ethics: critical-level utilitarian principles. *Econometrica*. 63. pp. 1303–20.

Broome, J. (1991) *Weighing Goods*. Cambridge, MA: Blackwell.

Broome, J. (2004) *Weighing Lives*. Oxford: Oxford University Press.

Buchak, L. (2016) Decision theory. In Hájek, A. and Hitchcock, C. (eds.). *The Oxford Handbook of Philosophy and Probability*. Oxford: Oxford University Press.

Diamond, P. (1967) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility: comment. *Journal of Political Economy*. 75. pp. 765–6.

Edgeworth, F. (1881) *Mathematical Psychics*. London: Kegan Paul.
Ellsberg, M. (1961) Risk, ambiguity and the Savage axioms. *Quarterly Journal of Economics*. 75. pp. 643–69.
Fishburn, P. (1970) *Utility Theory for Decision Making*. New York, NY: Wiley.
Fishburn, P. (1998) Stochastic utility. In Barberá, S., Hammond, P., and Seidl, C. (eds.) *Handbook of Utility Theory*. Vol. 1. Dordrecht: Kluwer.
Foot, P. (1985) Utilitarianism and the virtues. *Mind*. 94. pp. 196–209.
Gauthier, D. (1994) Assure and threaten. *Ethics*. 104. pp. 690–716.
Gilboa, I. (2009) *Theory of Decision under Uncertainty*. Cambridge: Cambridge University Press.
Hájek, A. (2012) Interpretations of probability. In Zalta, E. N. (ed.) *The Stanford Encyclopedia of Philosophy*. (Winter) [Online] Available from: http://plato.stanford.edu/archives/win2012/entries/probability-interpret.
Hájek, A. (2014) *Most Counterfactuals are False*. Manuscript.
Hammond, P. (1991) Interpersonal comparisons of utility: why and how they are and should be made. In Elster, J. and Roemer, J. (eds.). *Interpersonal Comparisons of Well-Being*. pp. 200–54. Cambridge: Cambridge University Press.
Hammond, P. (1998) Objective expected utility. In Barberá, S., Hammond, P. and Seidl, C. (eds.) *Handbook of Utility Theory*. Vol. 1. pp. 143–211. Dordrecht: Kluwer.
Harsanyi, J. (1953) Cardinal utility in welfare economics and in the theory of risk-taking. *Journal of Political Economy*. 61. pp. 434–5.
Harsanyi, J. (1955) Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*. 63. pp. 309–21.
Harsanyi, J. (1977a) Morality and the theory of rational behavior. *Social Research*. 44. pp. 623–56.
Harsanyi, J. (1977b) Nonlinear social welfare functions: a rejoinder to Professor Sen. In Butts, R. and Hintikka, J. (eds.). *Foundational Issues in the Special Sciences*. pp. 293–96. Dordrecht: Reidel.
Harsanyi, J. (1980) Rule utilitarianism, rights, obligations and the theory of rational behavior. *Theory and Decision*. 12. pp. 115–33.
Hausner, M. (1954) Multidimensional utilities. In Thrall, R., Coombs, C., and Davis, R. (eds.) *Decision Processes*. pp. 167–80. New York, NY: John Wiley & Sons.
Hoefer, C. (2007) The third way on objective probability: a sceptic's guide to objective chance. *Mind*. 116. pp. 549–96.
Jackson, F. (1991) Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics*. 101. pp. 461–82.
Jackson, F. and Smith, M. (2006) Absolutist moral theories and uncertainty. *Journal of Philosophy*. 103. pp. 267–83.
Jensen, K. (2012) Unacceptable risks and the continuity axiom. *Economics and Philosophy*. 28. pp. 31–42.
Joyce, J. (1999) *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
Kamm, F. (1996) *Morality, Mortality*. Vol. 2. New York, NY: Oxford University Press.
Kavka, G. (1983) The toxin puzzle. *Analysis*. 43. pp. 43–6.
Krantz, D., Luce, R. D., Suppes, P., and Tversky, A. (1971) *Foundations of Measurement*. Vol. 1. New York, NY: Academic Press.
Kreps, D. (1988) *Notes on the Theory of Choice. Underground Classics in Economics*. Boulder, CO: Westview Press.

Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.
Lewis, D. (1974) Radical interpretation. *Synthese*. 27. pp. 331–44.
Lewis, D. (1980) A subjectivist's guide to objective chance. In Jeffrey, R. (ed.). *Studies in Inductive Logic and Probability*. Vol. 2. pp. 83–132. Berkeley, CA: University of California Press.
Lewis, D. (1994) Humean supervenience debugged. *Mind*. 103. pp. 473–90.
McCarthy, D. (1998) Actions, beliefs and consequences. *Philosophical Studies*. 90. pp. 57–77.
McCarthy, D. (2008) Utilitarianism and prioritarianism II. *Economics and Philosophy*. 24. pp. 1–33.
McCarthy, D. (2013) Risk-free approaches to the priority view. *Erkenntnis*. 78. pp. 421–49.
McCarthy, D. (2015) Distributive equality. *Mind*. 124. pp. 1045–109.
McCarthy, D. (forthcoming a) The priority view. *Economics and Philosophy*.
McCarthy, D. (forthcoming b) *The Structure of Good*. Oxford: Oxford University Press.
McCarthy, D., Mikkola, K., and Thomas, T. (2016) Utilitarianism with and without expected utility. MPRA Paper No. 72578 https://mpra.ub.uni-muenchen.de/72578/.
McCarthy, D. and Thomas, T. (2016) Egalitarianism with risk. Manuscript.
McClennen, E. (1990) *Rationality and Dynamic Choice*. Cambridge University Press.
Mongin, P. (1995) Consistent Bayesian aggregation. *Journal of Economic Theory*. 66. pp. 313–51.
Myerson, R. (1981) Utilitarianism, egalitarianism, and the timing effect in social choice problems. *Econometrica*. 49. pp. 883–97.
Nagel, T. (1970) *The Possibility of Altruism*. Princeton, NJ: Princeton University Press.
Narens, L. (2007) *Introduction to the Theories of Measurement and Meaningfulness and the Use of Symmetry in Science*. Mahwah, NJ: Lawrence Erlbaum Associates.
Ng, Y. (1975) Bentham or Bergson? Finite sensibility, utility functions, and social welfare functions. *Review of Economic Studies*. 42. pp. 545–70.
Ok, E. (2007) *Real Analysis with Economic Applications*. Princeton, NJ: Princeton University Press.
Otsuka, M. and Voorhoeve, A. (2009) Why it matters that some are worse than others: an argument against the priority view. *Philosophy and Public Affairs*. 37. pp. 171–99.
Parfit, D. (1984) *Reasons and Persons*. Oxford: Clarendon Press.
Parfit, D. (2000) Equality or priority? In Clayton, M. and Williams, A. (eds.). *The Ideal of Equality*. pp. 81–125. Basingstoke: Macmillan.
Parfit, D. (2012) Another defense of the priority view. *Utilitas*. 24. pp. 399–440.
Rabinowicz, W. (2002) Prioritarianism for prospects. *Utilitas*. 14. pp. 2–21.
Ramsey, F. (1931) Truth and probability. In Ramsey, F. and Braithwaite, R. (ed.). *Foundations of Mathematics and other Essays*. pp. 83–132. London: Kegan, Paul, Trench, Trubner, & Co.
Rawls, J. (1971) *A Theory of Justice*. Cambridge, MA: Harvard University Press.
Rawls, J. (1982) Social unity and primary goods. In Sen, A. and Williams, B. (eds.) *Utilitarianism and Beyond*. Cambridge: Cambridge University Press.
Resnik, M. (1987) *Choices: An Introduction to Decision Theory*. Minneapolis, MN: University of Minnesota Press.
Roberts, F. (2009) *Measurement Theory*. Cambridge: Cambridge University Press.
Savage, L. (1954) *The Foundations of Statistics*. New York, NY: John Wiley.
Scanlon, T. (1982) Contractualism and utilitarianism. In Sen, A. and Williams, B. (eds.) *Utilitarianism and Beyond*. Cambridge, MA: Cambridge University Press.
Scheffler, S. (1982) *The Rejection of Consequentialism*. Oxford: Oxford University Press.

Schmidt, U. (2004) Alternatives to expected utility: formal theories. In Barberá, S., Hammond, P., and Seidl, C. (eds.) *Handbook of Utility Theory*. Vol. 2. pp. 757–837. Dordrecht: Kluwer.

Schwarz, W. (2016) Best system approaches to chance. In Hájek, A. and Hitchcock, C. (eds.) *The Oxford Handbook of Philosophy and Probability*. Oxford: Oxford University Press.

Sen, A. (1976) Welfare inequalities and Rawlsian axiomatics. *Theory and Decision*. 7. pp. 243–62.

Sen, A. (1985) Well-being, agency and freedom. *Journal of Philosophy*. 82. pp. 169–221.

Skyrms, B. (1990) *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

Sugden, R. (2004) Alternatives to expected utility: foundations. In Barberá, S., Hammond, P. and Seidl, C. (eds.). *Handbook of Utility Theory*. Vol. 2. pp. 685–755. Dordrecht: Kluwer.

Temkin, L. (2012) *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. Oxford: Oxford University Press.

Thomson, J. (1976) Killing, letting die, and the trolley problem. *The Monist*. 59. pp. 204–17.

Thomson, J. (1986) Imposing risks. In Parent, W. (ed.) *Rights, Restitution, and Risk*. Cambridge, MA: Harvard University Press.

Thomson, J. (2001) *Goodness and Advice*. Princeton, NJ: Princeton University Press.

Vickrey, W. (1960) Utility, strategy, and social decision rules. *The Quarterly Journal of Economics*. 74. pp. 507–35.

von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wakker, P. (2010) *Prospect Theory: For Risk and Ambiguity*. Cambridge: Cambridge University Press.

Weymark, J. (1991) A reconsideration of the Harsanyi-Sen debate on utilitarianism. In Elster, J. and Roemer, J. (eds.). *Interpersonal Comparisons of Well-Being*. Cambridge: Cambridge University Press.

Williams, B. (1973) A critique of utilitarianism. In Smart, J. and Williams, B. (eds.) *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Williamson, T. (1994) *Vagueness*. New York, NY: Routledge.

CHAPTER 34

PROBABILITY AND THE PHILOSOPHY OF RELIGION

PAUL BARTHA

34.1 Introduction

Probabilistic reasoning lies at the heart of many important arguments concerning the existence of God and the rationality of religious belief. This chapter discusses two of the most famous: the *fine-tuning argument* and *Pascal's Wager*. Both arguments are vigorously debated. Popular expositions abound, yet their careful formulation and assessment rely upon ever more sophisticated applications of the probability calculus and other formal tools.

This chapter focuses on how ideas in the philosophy of probability and decision theory shape current discussion of these two classic arguments. A secondary theme is that benefits flow in the reverse direction as well. The philosophy of religion, with its focus on the very large and the very improbable (often in combination), provides fertile ground for the development and testing of ideas in formal epistemology and decision theory.

Such connections are nothing new. There is a long history of fruitful links between developments in probability theory and the philosophy of religion. Pascal's Wager, which appeared in *Pensées* (Pascal 1670/1910), marked the birth of decision theory (Hacking 1972/1994). Bayes' Theorem was discovered by one erudite clergyman, Thomas Bayes, and published posthumously by another, Richard Price (Bayes 1763). In his introduction to Bayes' essay, Price announced an application: the mathematics would "confirm the argument taken from final causes for the existence of the Deity." He had in mind a form of the design argument, of which the fine-tuning argument is a contemporary descendant. Price also used Bayes' theorem to provide some of the earliest and most vigorous criticism of Hume's essay on miracles (Price 1767/1768).

Bayesian reasoning continues to play a prominent role in the philosophy of religion, but illumination comes from other sources as well: likelihoodism, non-standard (infinitesimal) probability measures, hyperreal utilities, and more. These diverse forms of "philosophical technology" help us to construct precise models of our assumptions, to clarify objections, and sometimes to identify an entirely new approach to a venerable problem.

In this connection, it is appropriate briefly to comment on a general concern that sometimes leads to discomfort with formal work in the philosophy of religion. The concern is that moving to a particular formal framework, even one as flexible as Bayesian epistemology, may lead to distorted representation of a philosophical problem. This can happen if the logical or mathematical assumptions implicit in the formal framework have a poor fit with the problem. An alternative form of this concern is that technical devices themselves can become the focus of our attention, distracting us from substantive issues in the philosophy of religion.

To this concern, we offer an equally general response. There is at least as much scope for distortion if we work at the level of intuition, without a clear probability model.[1] Perhaps the most effective response to the concern about distortion, however, is that the use of formal methods is consistent with keeping an open mind. Rather than committing ourselves to a single formal approach, the strategy adopted below is to use a variety of techniques to discuss ideas and arguments. As for the worry about distraction, we should remember that the application of sophisticated probabilistic ideas extends a long tradition in the philosophy of religion.

This section continues with a brief review of probabilistic concepts and their application to the philosophy of religion. Section 34.2 discusses the fine-tuning argument and three important objections. Section 34.3 is devoted to Pascal's Wager, with emphasis on novel techniques for representing and interpreting the argument. Section 34.4 concludes by summarizing ways in which the relationship between the philosophy of religion and the philosophy of probability is mutually beneficial.

34.1.1 Theistic Arguments and the Nature of God

Russell (2013) introduces a helpful distinction between two categories of theism, or belief in God. *Thin theism* is the belief that there exists "some 'supreme intelligence' that is the origin, creator and governor of this world . . . with no commitment to some further, more specific set of attributes." Such a belief is "theoretically empty and practically useless": whatever attributes God may have are totally beyond the range of human understanding. *Thick theism*, by contrast, is belief that "presupposes a richer set of attributes, such as infinity, omniscience, omnipotence, and moral perfection." Swinburne (2004: p. 7) offers an example: "there exists necessarily a person without a body (i.e. a spirit) who necessarily is eternal, perfectly free, omnipotent, omniscient, perfectly good, and the creator of all things." The truth of any version of thick theism has profound implications for human belief and action, so its probability is also of great significance.

Both of the theistic arguments to be considered below involve substantive assumptions about God's attributes. The fine-tuning argument aims at a conclusion about the probability that the universe is the result of intelligent design. In most versions, this requires the conception of God as a designer whose aims include the creation of a world of living creatures. Pascal's Wager concludes that a rational person should strive to believe in God. This argument requires the characterization of God as a being who can offer an infinite

[1] The "normalization objection" to the fine-tuning argument, in section 34.2.3, provides a good example.

reward, salvation, to human beings. Since our focus is on the structure of the two arguments, we do not provide any defence or criticism of these assumptions, but we do show how they are incorporated into the premises of the arguments.

34.1.2 Probabilistic Reasoning: A Quick Review

We shall be concerned primarily with subjective probabilities, i.e., *credences* or *degrees of belief*.[2] Bayesian epistemology provides the dominant philosophical framework for representing uncertainty and learning using such probabilities. The basic elements of Bayesian epistemology, reviewed elsewhere in this volume, are taken for granted.[3] Extensions of and alternatives to the Bayesian framework will be introduced as necessary in the course of the chapter.

By way of notation, we use $Pr(\cdot)$ and $Pr(\cdot/\cdot)$ to represent an agent's unconditional and conditional credences. By a probability model $< \Omega, F, Pr >$, we mean a set Ω of alternative possibilities (the *outcome space*), an algebra (or σ-algebra) F of subsets of Ω on which probabilities are defined, and a measure Pr that satisfies the standard probability axioms. We assume that where the agent has well-defined prior probabilities and appropriate conditional probabilities, she rationally updates via conditionalization.

Two constraints on prior probabilities are especially relevant to the philosophy of religion: *Regularity* and the *Principle of Indifference*. The most promising version of *Regularity* states that any doxastic possibility (any proposition that is logically compatible with the agent's beliefs) should receive positive probability.[4] Within the philosophy of religion, *Regularity* has to be applied with great care. For example, a key premise of Pascal's Wager is that the probability of God's existence is a positive real number. *Regularity* should not give us this premise for free! There is a different problem for the fine-tuning argument: there are too many hypotheses about the origin of the universe for each to have positive probability. Provided that *Regularity* is interpreted as compatible with *infinitesimal* probability assignments, however, we take it to be a reasonable guideline in assessing both arguments.

The *Principle of Indifference* (PI) states that if we have a family of mutually exclusive possibilities, and our evidence gives no more reason to believe any one of them more likely to be true than any other, then each possibility must be assigned equal prior probability (or equal conditional probability on the evidence).[5] Typically, PI is applied on the basis of an appeal to symmetry among the possible outcomes. As with *Regularity*, the *Principle of Indifference* has to be employed cautiously.[6] At this point, we simply register that PI plays a prominent role in our discussion of both the fine-tuning argument and Pascal's Wager.

[2] See Hájek (2011) for an introduction to the probability calculus and a review of different concepts of probability. See also the other chapters in this volume, especially (17) (Zynda 2016).

[3] In this volume, see chapters (17) (Zynda 2016), (18) (Sprenger 2016), (29) (Kotzen 2016), and (30) (Crupi and Tentori 2016).

[4] See Hájek 2013 for arguments that this and other versions of *Regularity* are indefensible.

[5] See Howson and Urbach 1993: p. 52. The Principle is generally attributed to Laplace (1820 VII: p. 181).

[6] See Gillies 2000 and van Fraassen 1989 for historical and philosophical discussion, including arguments that the principle is ultimately indefensible.

As a final preliminary point, we note that probabilistic arguments come in many different varieties. Three types will be especially important here.[7]

1. *Bayesian confirmational arguments* conclude that the evidence E raises or lowers the probability of some hypothesis H: $Pr(H/E) > Pr(H)$, or $Pr(H/E) < Pr(H)$. More precisely, confirmation is a three-place relation: E *confirms* (or *disconfirms*) H relative to background assumptions K if $Pr(H/E \,\&\, K) > Pr(H/K)$ (or $Pr(H/E \,\&\, K) < Pr(H/K)$).

2. *Likelihood arguments* conclude that the evidence E *favours* one hypothesis, H_1, over another, H_2. These arguments are based on likelihood comparisons, as summarized by the *Law of Likelihood*.

   (LL) Law of Likelihood.
   E favours H_1 over H_2 if and only if $Pr(E/H_1) > Pr(E/H_2)$.[8]

   Likelihood arguments require no assumptions about prior probabilities.

3. Finally, *pragmatic* (or *decision-theoretic*) arguments conclude, on the basis of subjective probabilities and information about the agent's preferences, that it is rational to take some course of action.

34.2 The Fine-tuning Argument

34.2.1 Introduction

According to the fine-tuning argument, the structure of our universe provides evidence for the existence of an intelligent designer. The starting point is the observation that the values of several fundamental physical constants lie within an extremely narrow range compatible with the existence of life.[9] Examples include the ratio of the masses of the electron and proton, and the fine-structure constant that characterizes the strength of electromagnetic interactions. If these values were just slightly different, living things could not exist.[10] In our universe, as Sober (2004) puts it, *the constants are right* for life. The probability that *the*

[7] For a more detailed discussion of probabilistic arguments in the philosophy of religion, see McGrew 2011 and Swinburne 2004.

[8] See Hacking 1965, Edwards 1972, Royall 1997 for discussion. Normally, H_1 and H_2 are mutually exclusive hypotheses. The likelihoods should incorporate relevant background assumptions (omitted here).

[9] The argument can also be based on fine tuning in the laws of nature, or in the initial conditions at the birth of the universe (Collins 2009: p. 202). In order to highlight the logical structure of the fine-tuning argument, we focus here on the values of the physical constants.

[10] Barrow 2001: p. 147. Similar examples abound; see Barrow and Tipler 1986, Leslie 1989; 1998, McMullin 1993, Ross 1998.

constants are right if their values have been set by chance is vanishingly small. By contrast, the probability is high if they have been set by an intelligent designer (God). This leads to the conclusion that the finely tuned physical constants are evidence for intelligent design (and for an intelligent designer).

The fine-tuning argument is a *cosmic design argument*. It should be carefully distinguished from the older *organismic design argument*, of which Paley (1802) provides the classic formulation. The starting point for the latter argument is the observation that organisms are delicately adapted to their environments. The probability of finding such delicate adaptations is much higher if they are the product of intelligent design than if they arose by chance; hence, observation favours the hypothesis of intelligent design. One familiar difficulty here is the availability of a powerful and well-confirmed alternative explanation for the adaptations, namely, evolutionary theory.[11] A second difficulty is to find plausible values (or ranges of values) for the relevant probabilities. The cosmic design argument strikes many people as promising because it appears to be free from both difficulties. There seems to be no possible evolutionary explanation for fine-tuning.[12] Furthermore, as we shall see, there is some hope for obtaining plausible values, or at least plausible bounds, for the probabilities in the fine-tuning argument.

The core idea behind the fine-tuning argument is expressed by a likelihood inequality. Let

$O \equiv$ The constants are right.
$H_{Ch} \equiv$ The universe was created by a chance process.
$H_{Des} \equiv$ The universe was created by intelligent design.

The likelihoods of interest are $Pr(O/H_{Des})$, the likelihood of the intelligent design hypothesis on the evidence, and $Pr(O/H_{Ch})$, the likelihood of chance on the same evidence. The likelihood inequality asserts:

(LI) $Pr(O/H_{Des}) > Pr(O/H_{Ch})$.

Typically, those who endorse the fine-tuning argument assert that this inequality is very large.

As Plantinga (2011) notes, there are basically three versions of the fine-tuning argument built around (LI). The first and most direct is a *likelihood* argument (Sober 2004; 2009). In this version, the probabilities in (LI) are understood to be objective. Since they can have no meaningful empirical basis, these likelihoods must be derived *a priori*. We then proceed directly from (LI) to the conclusion that the evidence objectively favours H_{Des} over H_{Ch}, citing the law of likelihood (LL) as justification.

Secondly, there is a *Bayesian* version of the argument (Swinburne 2004). The idea is that the observation that the constants are right incrementally confirms H_{Des},

[11] Sober 2004 argues that this difficulty does not affect the likelihood formulation of the design argument, but does undermine the Bayesian version.

[12] Smolin's cosmological natural selection theory (Smolin 1997) allows for a type of evolution of universes, but this is not related to the concerns of the fine-tuning argument.

(1a) $Pr(H_{Des}/O) > Pr(H_{Des})$,

and *disconfirms* H_{Ch},

(1b) $Pr(H_{Ch}/O) < Pr(H_{Ch})$.

The probabilities here are interpreted as subjective credences. In particular, the prior probabilities $Pr(H_{Des})$ and $Pr(H_{Ch})$ represent degrees of belief in design and chance before taking the evidence of fine-tuning into account.

Thirdly and most ambitiously, the fine-tuning argument can be formulated as an *inference to the best explanation* (Craig 2003, Dembski 1998; 2003, Colyvan et al. 2005). I shall not discuss this third approach.

Section 34.2.2 provides detailed formalizations of the likelihood and Bayesian versions of the fine-tuning argument. We then turn to three important objections. Section 34.2.3 discusses the *normalization objection*, which alleges that $Pr(O/H_{Ch})$ is *mathematically meaningless* (McGrew et al. 2001, Colyvan et al. 2005). If this objection succeeds, then (LI) is also meaningless and no version of the fine-tuning argument gets off the ground.

Section 34.2.4 discusses the objection that the fine-tuning argument is undermined by an *observation-selection effect*. This criticism was pressed early by Carter (1974/1989) and is prominent in the work of Sober (2004; 2009). The key idea is that likelihood reasoning is constrained by the *Weak Anthropic Principle*: "what we can expect to observe must be restricted by the conditions necessary for our presence as observers" (Carter 1974/1989). It comes as no surprise—and provides no support for intelligent design—that we observe constants fine-tuned for life; otherwise, we would not even be here to make the observation. Sober alleges that an observation-selection effect defeats the likelihood version of the fine-tuning argument because it makes the inequality (LI) unusable. Despite the prominence of this objection, its legitimacy is a matter of ongoing debate (Sober 2004, Weisberg 2005, Sober 2009). Furthermore, as Sober suggests, it leaves open the possibility of an alternative probabilistic argument which, I believe, may be interpreted as a Bayesian formulation.

Section 34.2.5 presents the *lottery objection*, which specifically targets the Bayesian version: fine-tuning provides no more confirmation for the design hypothesis than a winning lottery ticket provides for the hypothesis that the lottery was rigged. This objection has received little attention because it is masked by the two preceding objections. We don't see the problem unless we can get past them. Sections 34.2.5 and 34.2.6 show that even if we model the fine-tuning argument in a way that neutralizes those two objections, the lottery objection constitutes an important further challenge.

34.2.2 What is the Fine-tuning Argument?

A major attraction of the fine-tuning argument is that its basic concepts and assumptions appear amenable to a clear formulation. The argument has to do with *observed*, *life-permitting*, and *possible* values of certain physical parameters.[13] Since our primary concern is with the logical and mathematical structure of the argument, I make a few

[13] The content of these assumptions is discussed with great care in Collins (2009).

simplifying assumptions. The first is that there is only *one* relevant physical parameter, which we call k.[14] The second assumption is that the range of possible values for k is either the positive real numbers, $\mathbb{R}^+ = \{x/x > 0\}$, or some large bounded interval in $\mathbb{R}^+$.[15] The third assumption is that the observation O that *the constants are right* can be expressed by the statement $k \in J$, where J is a bounded interval of life-permitting values. That is, J contains all possible values of k compatible with the emergence of life.

34.2.2.1 *Likelihood formulation*

Given these simplifications, the *likelihood formulation* of the fine-tuning argument has three premises:

(2) O (*the constants are right*): $k \in J$. The Universe contains life.[16]
(3) If the constants were slightly different (if $k \notin J$), the Universe could not support life.
(4) $Pr(O/H_{Des})$ is high. That *the constants are right* is not surprising if the Universe is the product of intelligent design.

Premise (2) is obvious, and we shall simply accept (3). From (3) and the fact that the range of possible constants is either very large or all of $\mathbb{R}^+$:

(5) $Pr(O/H_{Ch})$ is low.

Hence, from (4) and (5):

(6) $Pr(O/H_{Des}) > Pr(O/H_{Ch})$.

This is just the likelihood inequality, (LI). Invoking the *Law of Likelihood* (LL), we obtain

(7) O (strongly) favours H_{Des} over H_{Ch}.

Clearly, the conclusion (7) rests entirely upon the comparison of the two likelihoods.

How do we obtain (4) and (5)? We need to be clear about our two hypotheses. Here is a proposal for H_{Des}: k is selected (randomly) from J, the set of life-permitting values. This makes (4) trivial; indeed, $Pr(O/H_{Des}) = 1$. And here are two proposals for H_{Ch}: k is selected by a uniform distribution over $\mathbb{R}^+$, or (if there is an upper bound) k is selected by a uniform distribution over some large interval. If this works (the mathematical difficulties are discussed in section 34.2.3), then (5) follows because $Pr(O/H_{Ch})$ is either very small or

[14] Although it seems impressive to point to several such parameters, the fine-tuning argument and the objections to it are identical in structure for any finite number of physical parameters (Colyvan et al. 2005).

[15] If logical possibility is the relevant modality, then the range of values is R^+ or perhaps all of R—see Colyvan et al. 2005. If some extended type of physical possibility is invoked, the range might be a bounded interval. We shall consider both alternatives below.

[16] We could state (2) as the stronger premise: we observe that the constants are right. What matters (especially in section 34.2.4) is that (2) be understood as observational evidence.

zero. By interpreting both H_{Des} and H_{Ch} as specific hypotheses about how k is selected, we have an *a priori* derivation for both likelihoods, as required.

Of course, in order for this to count as a cosmic design argument, we have to make some questionable additional assumptions. Let us say that a *selection hypothesis* corresponds to a subset S of possible values for k, together with a probability distribution over S. Any selection hypothesis, including random selection from J and random selection from the entire range of possible values, is consistent with either intelligent design or a mindless process. We cannot know, *a priori*, the intentions of a possible designer or the constraints on the possible physical processes that shape the universe. But let us be deliberately charitable to the fine-tuning argument by making two provisional assumptions that link selection to design. First, if k is selected from J, it can only be by intelligent design. Secondly, if k is selected randomly from the entire range of possible values, it can only be by a mindless process. The fine-tuning argument then becomes a comparison between two specific selection hypotheses, H_{Des} and H_{Ch}. It should be clear that this way of representing things is very favourable towards the fine-tuning argument.[17] We could easily call the two provisional assumptions into question. We shall not do so, however, because even if they are conceded, the fine-tuning argument faces severe problems.

34.2.2.2 *Bayesian formulation*

A Bayesian formulation of the fine-tuning argument begins with the same premises (2)–(4). By Bayes' Theorem,

(8a) $\Pr(H_{Des}/O) = \frac{\Pr(O/H_{Des})\Pr(H_{Des})}{\Pr(O)}$

and

(8b) $\mathbf{Pr}(H_{Ch}/O) = \frac{\mathbf{Pr}(O/H_{Ch})\mathbf{Pr}(H_{Ch})}{\mathbf{Pr}(O)}$.

The conclusion of the Bayesian argument (as stated above) is that the evidence confirms H_{Des} and disconfirms H_{Ch}:

(1a) $Pr(H_{Des}/O) > Pr(H_{Des})$

and

(1b) $Pr(H_{Ch}/O) < Pr(H_{Ch})$.

Referring to (8a) and (8b), we see that these conclusions follow if we establish

(9a) $Pr(O) < Pr(O/H_{Des})$

[17] As Sober 2004 points out, Premise (4) is not obvious if H_{Des} represents a general hypothesis of design. For discussion, see Swinburne 2004 and Collins 2009. Sober also maintains that if we entertain a particular hypothesis of design for life-permitting values, then we should entertain particular chance hypotheses including mindless selection from the interval of life-permitting values.

and

(9b) $Pr(O) > Pr(O/H_{Ch})$,

respectively. In order to derive (9a) and (9b), we have to calculate

$$Pr(O) = \Sigma_H Pr(O/H)Pr(H),$$

averaging over all possible selection hypotheses H.

We can appreciate why the confirmation version of the fine-tuning argument is much more complicated than the likelihood version. Although our attention is focused on just two hypotheses, in order for the argument to succeed, we need to define a full probability model. We first have to specify a set of *potential selection hypotheses*, $\mathcal{H}$. Each such hypothesis, as noted earlier, corresponds to a subset S of $\mathbb{R}^+$ of possible values for k together with a probability distribution over S. Taking the range of possible values for k to be $\mathbb{R}^+$, the space Ω of *possible outcomes* is $\mathbb{R}^+ \times \mathcal{H}$:

$$\Omega = \{ < k, H > / k \in \mathbb{R}^+ \text{ and } H \in \mathcal{H} \}.$$

We then have to define a credence function Pr, a joint probability distribution on certain subsets of Ω.[18] That is, we have to define:

(a) A distribution Pr over $\mathcal{H}$, i.e., over the selection hypotheses.
(b) Conditional probabilities $Pr(k \in I/H)$ for each H in $\mathcal{H}$ and for certain subsets I of positive real numbers.

Yet there is almost no limit to what counts as a possible selection hypothesis.

This leads to a fundamental point: in order to keep things manageable, any Bayesian probability model must exclude a large class of possible selection hypotheses. But how can we judge that some selection hypothesis would never be considered by an unknown intelligent designer, or would be incompatible with any possible physical process? As Sober puts it, "the assumption that God can do anything is part of the problem" (2004: p. 41).

Despite the uncertainty about which possibilities to include, I suggest that a Bayesian probability model can still be informative, provided that it meets three conditions. First, it should be *comprehensive*; it should include all types of hypotheses that have been seriously entertained. Secondly, the hypotheses H in $\mathcal{H}$ should be *specific* enough that the likelihoods $Pr(O/H)$ are well-defined. Thirdly, and most controversially, the set $\mathcal{H}$ and the prior distribution Pr should be *unbiased*: we should have $Pr(k \in I) = Pr(k \in I')$ whenever I and I' are finite intervals of equal size within the range of possible values for k.[19] It would be significant if a version of the fine-tuning argument could succeed under these three constraints.

As a first step towards laying out such a model, we do well to confine our attention to hypotheses that take the form of *uniform distributions over an interval*, or over the

[18] At minimum, Pr will be defined on an algebra of subsets.

[19] Of course, a Bayesian can adopt any coherent set of prior probabilities. The requirement that the distribution should be unbiased is intended as a plausibility constraint. It will be scrutinized below.

complement of an interval. That is, if I is any interval in $\mathbb{R}^+$, consider selection hypotheses of two types:

- SEL_I: the value of k is selected via a uniform probability distribution over I (values outside I are excluded).
- $SEL_{\bar{I}}$: the value of k is selected via a uniform probability distribution over $\bar{I} = \{x \in \mathbb{R}^+ / x \notin I\}$ (values within I are excluded).

The interval I can be an ordinary bounded interval in $\mathbb{R}^+$ or an infinite interval.[20] The set of selection hypotheses is thus

$$\mathcal{H} = \{SEL_I \text{ and } SEL_{\bar{I}}/I \text{ is any interval in } \mathbb{R}^+\}.$$

Although this omits many possibilities, $\mathcal{H}$ is reasonably comprehensive: it includes most selection hypotheses ordinarily considered in cosmic design arguments. In particular, note that $H_{Ch} = SEL_{(0,\infty)}$ represents a uniform distribution over $\mathbb{R}^+$. Also, $H_{Des} = SEL_J$ stands for a uniform distribution over the interval J of life-permitting values of k.

The final step in defining our probability model is to stipulate a distribution over $\mathcal{H}$. We postpone that task because, already, mathematical red flags should be going up. What is a uniform distribution over $\mathbb{R}^+$, as required for H_{Ch}? More generally, what is a uniform distribution over I when the interval is unbounded? This worry is the normalization objection, to which we now turn.

34.2.3 The Normalization Objection

Suppose that the range of possible values for k is the positive real numbers, $\mathbb{R}^+$. The chance distribution is supposed to be uniform in the following sense:

If I and I' are two intervals of equal length, then $Pr(k \in I/H_{Ch}) = Pr(k \in I'/H_{Ch})$.

However, there is no uniform probability distribution over the positive real numbers. The hypothesis H_{Ch} is mathematical nonsense, $Pr(O/H_{Ch})$ is meaningless, and therefore the likelihood inequality (LI) is meaningless. The conditional probabilities in the Bayesian version are similarly meaningless whenever they involve uniform distribution over an unbounded interval. This is the *normalization objection*. It was explained clearly by McGrew et al. (2001) and further developed by Colyvan et al. (2005).

The difficulty is a consequence of countable additivity. Suppose that μ is a uniform probability measure on $\mathbb{R}^+$, i.e., μ assigns the same measure to any two finite intervals of equal length. Let $\mu((0, 1]) = c$ be the measure assigned to the interval $(0, 1] = \{x/0 < x \leq 1\}$. By uniformity, $\mu((n, n+1]) = c$ for $n = 1, 2, \ldots$. By countable additivity,

$$\mu(\mathbb{R}^+) = \Sigma_n \mu((n, n+1]) = c + c + \ldots.$$

[20] Here, we assume that there is no upper bound on the possible value of k. We consider imposing a bound in section 34.2.5.

If $c = 0$, the sum is 0; if $c > 0$, the sum is infinite. Since any probability measure is normalized, we must also have $\mu(\mathbb{R}^+) = 1$. This implies that $c + c + \ldots = 1$. Whether $c = 0$ or $c > 0$, we have a contradiction. There can be no such probability measure.

In the likelihood version of the fine-tuning argument, trouble arises at step (5), which asserts that if J is the interval of life-permitting values, then

$$Pr(k \in J/H_{Ch}) = \varepsilon,$$

for some small value ε. This is supposed to follow from (3). But it doesn't follow, since there is no way to make sense of H_{Ch} in terms of a conventional probability distribution. In the Bayesian version, there is a parallel problem for step (9b). Consequently, neither the likelihood argument nor the Bayesian argument is valid.

One possible response to this objection is to give up countable additivity. We settle for a finitely additive uniform measure, μ, on $\mathbb{R}^+$. If we do this, we can assign equal measure to any two intervals of equal length without contradiction. In fact, in this case, we must have $\mu(A) = 0$ for any bounded interval A, so $Pr(k \in A/H_{Ch}) = 0$. As noted by McGrew et al. (2001) and Colyvan et al. (2005), however, this proposal runs contrary to the spirit of the fine-tuning argument. For one thing, the argument no longer depends on fine-tuning: it works just as well no matter how large the range of life-permitting values, so long as that range is finite. There is a further problem: *any value* for k, even one that is not life-permitting (if it could be observed), favours any hypothesis of selection for a bounded interval that contains k over the hypothesis of chance. The use of a finitely additive measure stacks the deck against the chance hypothesis.[21] So the proposal is a non-starter.

Collins (2009) endorses the above "coarse-tuning" strategy (in the terminology of McGrew et al. 2001). He writes:

> Assume that the fine-tuning argument would have probative force if the comparison range were finite … Now imagine increasing the width of this comparison range while keeping it finite. Clearly, the more W_R [the comparison range] increases, the stronger the fine-tuning argument gets … Accordingly, if we deny that [the argument] has probative force *because* W_R is purportedly infinite, we must draw the counterintuitive consequence that although the fine-tuning argument gets stronger and stronger as W_R grows, magically when W_R becomes actually infinite, the fine-tuning argument loses all probative force.
>
> (pp. 251–2)

The finite case is a uniform measure over a bounded comparison range, $W_R = [0, N]$. The limiting case is a finitely additive uniform measure over the comparison range $W_R = \mathbb{R}^+$. Collins assumes that those raising the normalization objection concede that the fine-tuning argument goes through when the comparison range is bounded, and even becomes stronger as that range increases in size. Since discontinuity at the limit is counterintuitive, Collins reasons, we should conclude that the fine-tuning argument has probative force even when the comparison range is infinite. This is an unconvincing argument. Philosophical discomfort with discontinuity should not trump a mathematically-based objection.

[21] This is related to the non-conglomerability of finitely additive measures; see Kadane et al. 1996. For Bayesian versions of this argument, the posterior probability of H_{Ch} drops to 0 no matter what value of k is observed.

There are three possible responses to the normalization objection. The first is to consider versions of the fine-tuning argument that impose a cap or upper bound on the comparison range: $k \in [0, N]$ for some N. As pointed out by both McGrew et al. (2001) and Colyvan et al. (2005), there seems to be no non-arbitrary way to impose such a cap, but we shall consider this approach in section 34.2.5. A second response, just as vulnerable to the charge of arbitrariness, is to assume that the chance distribution $Pr(\cdot/H_{Ch})$, is not uniform but favours small values for k over large ones. A third response, outlined in section 34.2.6, is to employ a non-standard probability measure. Before considering these ideas, we turn to the second major objection to the fine-tuning argument.

34.2.4 The Observer-Selection Effect Objection

An observer-selection effect (OSE) occurs in the observation of O when the observation procedure itself affects one or more of the relevant conditional probabilities for observing O.[22] The effect is clearest when an entire class of possible observations is ruled out. Sober (2004) illustrates the idea with an example due to Eddington (1939). Suppose that you go fishing with a net that has 10-inch holes in the mesh. You observe O: all the fish that you catch are more than 10 inches long. Let H_1 be the hypothesis that all the fish in the lake are more than 10 inches long and H_2 the hypothesis that half the fish in the lake are more than 10 inches long. Since $Pr(O/H_1) = 1$ and $Pr(O/H_2) = 1/2^n$, we have the likelihood inequality:

(10) $Pr(O/H_1) > Pr(O/H_2)$.

You might think that the observation O favours H_1, the hypothesis with the greatest likelihood. But this inference neglects the observation procedure. Let P represent the procedure: you use a net that catches only fish more than 10 inches long. Given P, no possible observation can discriminate between H_1 and H_2. The relevant likelihoods, which take P into account, are therefore $Pr(O/H_1 \,\&\, P)$ and $Pr(O/H_2 \,\&\, P)$. Since

(11) $Pr(O/H_1 \,\&\, P) = Pr(O/H_2 \,\&\, P) = 1$,

the evidence O does not favour either hypothesis. This result agrees with intuition.

In general, whenever the observation procedure P influences the relevant likelihoods for observing O, you cannot use (10) to infer that the evidence O favours H_1 over H_2. Instead, the relevant likelihood comparison is whether

(12) $Pr(O/H_1 \,\&\, P) > Pr(O/H_2 \,\&\, P)$.

How does this apply to the fine-tuning argument? In this case, the observation procedure P requires that humans exist (to make observations). This can only happen if the constants are right. So, conditional upon P, if we check the values of the constants then we should expect to observe that they are right, whether the universe is the result of design or mindless chance. That is, letting O now stand for *the constants are right*,

[22] See Bostrom 2002 for extended discussion.

(13) $Pr(O/H_{Des} \,\&\, P) = Pr(O/H_{Ch} \,\&\, P) = 1$,

The likelihood inequality (6), which asserts that $Pr(O/H_{Des}) > Pr(O/H_{Ch})$, becomes irrelevant. The observation that the constants are right favours neither H_{Des} nor H_{Ch}. Sober (2004; 2009) concludes that the observation-selection effect defeats the likelihood version of the fine-tuning argument.

Sober's argument has led to vigorous debate. Three types of objection are salient. First, critics put forward parallel cases where likelihood reasoning seems unimpeachable, but would apparently be defeated by an observation-selection effect if Sober's reasoning were correct. Suppose that the prisoner facing the firing squad entertains two hypotheses: either the soldiers intend to spare his life or the soldiers are shooting to kill. Doesn't his observation that he is alive, a short time after the shots ring out, count as evidence favouring the former hypothesis over the latter—despite the fact that he has to be alive to make the observation (Leslie 1989, Swinburne 1990, Weisberg 2005)?

Secondly: following Weisberg (2005), we might think to identify a fallacy in Sober's reasoning. Considering the set-up of the fine-tuning argument, Weisberg alleges that the observation procedure P entails only the following conditional: *if* I observe whether the universe is fine-tuned, *then* I will find that it is (2005: p. 819). This conditional does not entail that I exist, nor does it entail O, that the constants are right. Since P does not entail O, (13) is false. In fact, the conditional appears to be statistically irrelevant to both likelihoods. This suggests that it makes no difference when we conditionalize on P, and (12) is correct (putting H_{Des} for H_1 and H_{Ch} for H_2).

Thirdly: Sober grants that a "probability argument" can be offered in place of the likelihood version of the fine-tuning argument, and that it faces no observation-selection effect. Similarly, he asserts that a probability argument may be employed by the firing-squad victim. Shortly, I shall suggest that this is a Bayesian argument. If the likelihood argument is defeated by an observation-selection effect, however, and the Bayesian argument must also factor the observation procedure into the likelihoods, how can the Bayesian argument succeed?

Sober's best response to these objections is in Sober (2009). Let's start with the second objection. Sober insists that the observation procedure, P, includes one's existence (or the existence of one's ancestors) at a time t_2 prior to the observation of the constants at t_3. He writes:

> Notice that Weisberg does not conditionalize on the fact that I am alive at t_2. I do so because I think this simple fact is part of the process leading to my observation at t_3. It is just as much part of the process as the choice of a net with which to fish.
>
> (2009: p. 84)

The observation process includes facts from the causal history preceding the observation that establish the relevant conditional probabilities. In the case of the fine-tuning argument, this gives us (13).

Although I think that Sober's reasoning is essentially correct, we need further clarification of what we mean by the observation process. As Sober notes, we don't want the entire causal history leading up to an observation to be part of the observation process. Otherwise, all likelihoods would be 1 or close to 1, and observation-selection effects would be

ubiquitous. The observation process includes everything that establishes the possible values and conditional probabilities of the observable random variables.

Sober's fundamental insight is that the fine-tuning argument is a special case of a type of situation where there is more than one possible observation procedure. As a simple illustration, he asks us to consider a variant of Eddington's fishing story in which there is an initial selection between a net with 10-inch holes and a net with 2-inch holes. In this story, there are two possible "pure" observation procedures, P_1 (10-inch holes) and P_2 (2-inch holes), and there is also a "mixed" observation procedure $P = \alpha\ P_1 + (1 - \alpha)P_2$ if it is a matter of chance which net is used (where $\alpha = Pr(P_1)$). Suppose, as before, that you observe (O) only fish longer than 10 inches in the net. As before, you want to determine whether the evidence favours H_1 (all fish are longer than 10 inches) or H_2 (half the fish are longer than 10 inches). If you happen to know that the large net was used (you saw the holes!), the fact that the net was selected by chance is irrelevant. In your likelihood comparison, you should use P_1 rather than P. Since $Pr(O/H_1\ \&\ P_1) = Pr(O/H_2\ \&\ P_1) = 1$, the evidence favours neither hypothesis. It is not that your causal history, right up to the moment of observation, is part of P_1. P_1 includes only the history up to the moment when the net is selected. Your recent history does, however, play an indirect role because it makes you certain that the large net was used.

The situation for the fine-tuning argument is analogous in that once again, there are two possible 'observation procedures', at least in a formal sense. The first is a normal procedure P_1 with human observers. The second, P_2, is a degenerate or *null* observation procedure with no observers. Unlike the Eddington variant where the selection of observation procedure is independent of the hypotheses H_1 and H_2, the selection of procedure P_1 or P_2 is statistically linked to H_{Des} and H_{Ch}.[23] Nevertheless, we know that P_1 rather than P_2 is the actual observation procedure. So we should use P_1 in our likelihood comparison. Since P_1 implies O (*the constants are right*), both likelihoods are 1 and we obtain the familiar result: the observation O favours neither H_{Des} nor H_{Ch}. This accounts for the observation-selection effect.

In order to clarify this point further and to address our third concern—why observation-selection is *not* a problem for the Bayesian analysis—consider a new example that is simpler than either the firing-squad or the fine-tuning case.

> *Sleeping Fred*. Fred goes to sleep. A fair coin is then tossed twice at time t_1.[24] If at least one toss comes up *Heads*, then Fred is awakened at t_2. If instead the result is *Tails-Tails*, then Fred remains asleep (until some much later time). Fred learns all of this before he goes to sleep. Now suppose that Fred has been awakened at t_2 and observes at t_3 that he is awake and has been awake since t_2. Call this observation O. It seems that Fred has evidence that favours the hypothesis *Heads-on-first*: the result of the first toss was *Heads*.

The Bayesian analysis of this case appears to be very simple. Fred relies upon the likelihood inequality:

(14) $1 = \Pr(\text{O/Heads-on-first}) > \Pr(\text{O/Tails-on-first}) = ½.$

[23] As Sober suggests, we could restore the analogy by changing Eddington's example to make the chance selection of the large or small net vary with the composition of fish in the lake.

[24] If we like, we can modify the example by assuming a coin of unknown (though not extreme) bias.

Using Bayes' Theorem, Fred computes posterior probabilities: $Pr(\textit{Heads-on-first}/O) = 2/3$ and $Pr(\textit{Tails-on-first}/O) = 1/3$. The observation O confirms *Heads-on-first* and disconfirms *Tails-on-first*. The same reasoning could be used by an observer, say Susan, who is watching the process and can observe whether Fred is awake or asleep at t_2.

If Sober is right, however, it seems that Fred cannot use (14) in a likelihood argument because there is an observation-selection effect. Fred can only make observations if he is awake. The observation procedure P thus includes the fact that Fred is awake at t_2, and hence $Pr(O/P) = 1$. Indeed, Fred is certain that P is the actual observation procedure, rather than the null procedure where he stays asleep. Since

(15) Pr(O/Heads-on-first & P) = Pr(O/Tails-on-first & P) = 1,

likelihood reasoning favours neither *Heads-on-first* nor *Tails-on-first*.

At this point, we see an interesting difference between likelihoodists and Bayesians. Royall (2004: p. 138) notes that the likelihood paradigm "requires probability models for the observable random variables only." The only observable random variable here is O, and the only likelihood comparison available to Fred is (15). As a result, there is no favouring relation. For the Bayesian, by contrast, the probability model includes not just observable random variables, but also random variables P for the observation process (P = Fred is awake at t_2, $\sim P$ = Fred is asleep at t_2) and H for the hypothesis (H_1 = *Heads-on-first*, H_2 = *Tails-on-first*). Here is the Bayesian computation again, this time factoring in the observation procedure:

$$\begin{aligned}\Pr(H_1/O\ \&\ P) &= \frac{\Pr(O/H_1\ \&\ P)\Pr(H_1/P)}{\Pr(O/P)}\\ &= \Pr(H_1/P)\\ &= \frac{\Pr(P/H_1)\Pr(H_1)}{\Pr(P)}\\ &= 2/3,\end{aligned}$$

since $Pr(P/H_1) = 1$ and $Pr(P) = Pr(P/H_1)Pr(H_1) + Pr(P/H_2)Pr(H_2) = \frac{3}{4}$.
Similarly,

$$Pr(H_2/O\ \&\ P) = 1/3.$$

In the Bayesian analysis, the observation procedure factors out. In essence, we replace the uninformative likelihood comparison (15) with the informative comparison

(16) $Pr(P/H_1) > Pr(P/H_2)$.

This comparison is possible because we have a full probability model. The observation-selection effect disappears; indeed, the Bayesian analysis is the same for Fred as for our outside observer, Susan. Of course, as the computation of $Pr(P)$ shows, the Bayesian needs more assumptions than the likelihoodist, including well-defined likelihoods for the full range of possible hypotheses.

But isn't the likelihood comparison (16) also available to the likelihoodist? Sober briefly considers the analogous likelihood comparison for the fine-tuning argument:

(17) Pr(I exist/Intelligent Design) > Pr(I exist/Chance).

Here, we let "I exist" stand for the observation process. Sober notes that an argument for Intelligent Design based on the observation that I exist is different from the fine-tuning argument. He also raises questions that cast doubt on (17). But he omits the most fundamental problem. In light of his position that, in any likelihood argument, the likelihoods being compared must both take into account (conditionalize upon) the observation procedure, (17) is simply unavailable to us, just as (16) is unavailable to Fred. Likelihoodists cannot escape the tight circle that always brings them back to the observation procedure. Sober's doubts about (17), therefore, are relevant to a Bayesian formulation of the fine-tuning argument, but not to the likelihood formulation.

Finally, what about the criticism that Sober's analysis of the observation-selection effect has counterintuitive consequences in cases such as the firing squad? The firing squad makes its decision at t_1, shots ring out at t_2, and the prisoner finds himself still alive at t_3. Does the observation that he is alive at t_3 favour the hypothesis that the squad intended to spare him rather than kill him? In 2004, Sober thinks that an observation-selection effect blocks the likelihood reasoning. In 2009, he changes his mind: the firing-squad case is not analogous to the fine-tuning case. Conditionalizing on the observation process does not make all likelihoods equal to 1. Whether the soldiers shoot to kill or not, the prisoner is alive when they fire and while the bullets are in flight. There is only one possible observation process here, and it does not entail that the prisoner is alive at t_3.

There is room for doubt about this resolution of the firing-squad case. It requires that we think of the prisoner's observation as a process that occurs (or that begins to occur) shortly after t_2 and prior to t_3. This observation process has two possible results, one that is truncated abruptly shortly after t_2 (if the prisoner dies) and one that continues on to t_3 (if he lives). This changes the problem slightly from the original formulation: the evidence for a friendly firing squad is extended over time, rather than consisting simply of the prisoner's observation that he is alive at t_3. If we grant this change, then likelihoodist reasoning can successfully handle the firing-squad case.

There are, however, other counterintuitive consequences of Sober's position. Consider an extreme version of *Sleeping Fred*: the probability that he is awakened if the first toss is *Tails* is $1/N$ rather than ½, where N is astronomically large. Suppose, in fact, that there is only one coin toss. On *Heads*, Fred is awakened, but on *Tails*, Fred remains asleep unless there is an earthquake within five minutes of the toss that causes the coin to flip from *Tails* to *Heads*.[25] The observation selection effect exists just the same and no likelihood argument available to Fred can discriminate between *Heads-on-first* and *Tails-on-first*. For such cases, it seems, the likelihood approach does not reflect our ordinary concept of evidence. We should recognize, of course, that likelihoodists such as Sober have no problem with using Bayesian reasoning when objective information about prior probabilities is available (Sober 2002), as they are for Fred. The point of the preceding argument is not that Sober would

[25] My thanks to Alexander Pruss for this example.

have to embrace an absurd conclusion, but only that reaching the right conclusion in this case requires a Bayesian, rather than a likelihoodist, concept of evidence.

To summarize: an observation-selection effect defeats the likelihood argument for the fine-tuning argument. Given the actual observation procedure, the relevant observation (*the constants are right*) cannot discriminate between the two hypotheses under consideration. However, the door is still open to developing a Bayesian version of the fine-tuning argument, provided that we can avoid the normalization objection. Our next task is to show how this might be done.

34.2.5 Finite Models and the Lottery Objection

This section examines two finite Bayesian models for the fine-tuning argument. The first allows only finitely many integer values for k, i.e., $k \in \{1, 2, \ldots, N\}$. The second imposes an upper bound or "cap" on the set of positive real values, i.e., $k \in [0, N]$ for some N. The use of Bayesian models avoids the observation-selection objection; the use of finite models avoids the normalization objection. In both cases, however, the fine-tuning argument still runs into difficulty.

Suppose first that there are N possible values for k: we must have $k = 1$ or $k = 2$ or ... or $k = N$. Suppose further that exactly one of these values, $k = k^*$, is consistent with the emergence of life, and that the observed evidence O is that $k = k^*$. It is easy to define a probability model that meets the three conditions stated in section 2.2. For any subset S of $\{1, \ldots, N\}$, let SEL_S be the hypothesis that the value of k is selected via a uniform distribution on S. If S contains n elements, then

$$\Pr(k = i/SEL_S) = \begin{cases} 1/n, & i \in S \\ 0, & i \notin S \end{cases}.$$

For the set of potential selection hypotheses, take

$$\mathcal{H} = \{SEL_S/S \text{ is any subset of } \{1, \ldots, N\}\}.$$

In particular, $H_{Des} = SEL_{\{k^*\}}$ and $H_{Ch} = SEL_{\{1,\ldots,N\}}$.

To complete the definition of our probability model, we provide an unbiased initial credence function over the set $\mathcal{H}$. We should, however, have some freedom in assigning prior probabilities for H_{Ch} and H_{Des}. Let D_n represent the family of all selection hypotheses for subsets of size n:

$$D_n = \{SEL_S/S \text{ has } n \text{ elements}\}.$$

For each $1 \leq n \leq N$, D_n is a family of $\binom{N}{n}$ hypotheses. Note that D_N is the singleton set $\{H_{Ch}\}$. We can specify whatever prior probabilities we like for $D_1, \ldots, D_N$, but we insist on a uniform conditional probability over hypotheses within each D_n. That is, if $Pr(D_n) = \alpha_n$, we require that

$$\Sigma\alpha_n = 1$$

and that for any two subsets J and J' of $\{1, \ldots, N\}$ with n members,

$$Pr(J) = Pr(J') = \alpha_n / \binom{N}{n}.$$

We refer to this as the *lottery model*. It is the model that we adopt for a lottery with N tickets if we think that the lottery might be rigged to select the winner uniformly from within some favoured group of n tickets, but we lack any prior evidence that favours any such group.

What happens when we conditionalize on the evidence O that $k = k^*$? The probability for each D_n remains unchanged (see *Appendix*):

(18) $Pr(D_n/O) = Pr(D_n) = \alpha_n$.

Not only is there no confirmation for $H_{Ch} = D_N$, but also no adjustment in our probability for each *family* D_n of hypotheses. We do have confirmation for design:

(19) $Pr(H_{Des}/O) = \alpha_1 > \alpha_1/N = Pr(H_{Des})$.

In fact, the observation that $k = k^*$ confirms every hypothesis SEL_J such that $k^* \in J$. But the probability increase arises entirely from a re-distribution of the probability within each family D_n, to account for selection hypotheses that have been ruled out. In particular, there is no disconfirmation of H_{Ch}:

(20) $Pr(H_{Ch}/O) = Pr(H_{Ch}) = \alpha_N$.

Relative to this prior probability distribution, the Bayesian argument fails because the conclusion (1b) is false. The observation of fine-tuning ($k = k^*$) provides *no meaningful confirmation* for design.

The existence of this model shows that the Bayesian version of the fine-tuning argument is invalid, even if we can overcome the normalization objection. Call this the *lottery objection* because of the analogous case of a possibly rigged lottery. The analogy is helpful. Although we suspect corruption, we start with an unbiased distribution. When some ticket k^* is the winner, this confirms hypotheses compatible with that result, at the expense of incompatible hypotheses. But there is no change to the probability that the lottery was conducted fairly.

We obtain analogous results if we shift from a discrete model to a version where the possible values of k are non-negative real numbers with an upper bound N, i.e., $0 \leq k \leq N$. We identify a sub-interval J as the set of life-permitting values, and we consider the class $\mathcal{H}$ of all hypotheses SEL_I and $SEL_{\bar{I}}$, as explained in section 34.2.2, where I is a sub-interval of $[0, N]$. We define families D_r consisting of all hypotheses SEL_I and $SEL_{\bar{I}}$ such that the measure of I or $\bar{I}$ is r. We define an unbiased credence function Pr that is uniform over hypotheses within each D_r. Just as before, if O is the evidence that $k \in J$, then $Pr(D_r/O) = Pr(D_r)$. We see that O makes no difference to the probability of the chance hypothesis; once again, the Bayesian version of the fine-tuning argument fails. In section 34.2.3, we noted that philosophers have objected to the arbitrariness of imposing a cap on possible values of the fundamental constants. We now see that even if we impose such a cap and successfully avoid the normalization objection, the fine-tuning argument falls to the lottery objection.

The above arguments rely on the assumption that our prior probability distribution is unbiased: it does not favour the selection of an interval of life-permitting values over any other interval of equal measure. Might we legitimately reject that assumption? To be sure, a Bayesian can adopt any coherent set of prior probabilities. Nevertheless, there are a number of considerations in favour of unbiased priors.

First and foremost: the fine-tuning argument should start on neutral ground. Prior to any evidence about the actual values of the constants, there is good reason to assign equal weight to selection of subsets of equal size. If the fine-tuning argument can succeed only by assuming a prior bias in favour of life-permitting values, that weakens the force of the argument considerably.

Secondly: although we cannot rule out the possibility of an *a priori* argument for biased priors, it is difficult to motivate the departure from neutrality. In the analogous case of a lottery, experience provides good reasons to reject an unbiased prior distribution. We are justified, for example, in assigning a slightly elevated prior probability to the proposition that the winning ticket is held by an associate or employee of a convenience store. When defining prior probabilities for the fine-tuning argument, by contrast, we have no experiential basis on which to draw.

Finally: even if the above points are disputed, the arguments of this section establish that the fine-tuning argument is invalid, unless one can demonstrate that we are *required* to have prior credences biased in favour of life-permitting values for the constants.

34.2.6 A Dilemma for the Fine-tuning Argument

Let us return to the version of the fine-tuning argument in which the range of possible values for k is the positive real numbers, $\mathbb{R}^+$. We sketch a way to circumvent the normalization objection by using a non-standard probability measure. If we use such a measure, however, then the argument becomes vulnerable to an analogue of the lottery objection.

We start with a brief explanation of how we can define a uniform measure on $\mathbb{R}^+$, using non-standard probabilities.[26] An *infinitesimal* ε is a number that is positive but smaller than any positive real number. Write $\mathbb{R}(\varepsilon)$ for the smallest closed field containing ε and $\mathbb{R}$. We want to define a non-standard probability measure on the family $\mathcal{F}$ of finite unions of intervals in $\mathbb{R}^+$. A *non-standard probability measure* ν on $(\mathbb{R}^+, \mathcal{F})$ with range $\mathbb{R}(\varepsilon)$ assigns a value in $\mathbb{R}(\varepsilon)$ to each interval I in F such that three properties are satisfied:

(NS1) $\nu(\mathbb{R}^+) = 1$;
(NS2) $0 \leq \nu(I) \leq 1$ for each interval I; and
(NS3) ν is *finitely additive*: $\nu(I \cup J) = \nu(I) + \nu(J)$ if I and J are disjoint intervals.

Here is one way to define a non-standard measure that is uniform.[27] If I is a finite interval, write $m(I)$ for the Lebesgue measure (or length) of I; if I is an infinite interval, then its

[26] For an introduction to non-standard analysis and hyperreal numbers, the reader can consult Hurd and Loeb 1985 or Goldblatt 1998.

[27] I follow the lead of Halpern 2010, who uses a similar technique to define a uniform measure on the natural numbers.

complement $\bar{I}$ is a finite interval with a well-defined length $m(\bar{I})$. Define

$$\nu(I) = \varepsilon \cdot m(I) \qquad \text{if } I \text{ is a finite interval, and}$$
$$\nu(I) = 1 - \varepsilon \cdot m(\bar{I}) \qquad \text{if } I \text{ is an infinite interval.}$$

ν is well-defined for unions of intervals by finite additivity. Furthermore, ν is uniform because it assigns the same measure to any two finite intervals of equal length.

Although ν is *not* countably additive, it does distinguish between finite intervals of different size, just as Lebesgue measure does. In fact, $\nu(I)/\nu(J) = m(I) / m(J)$ for any two finite intervals I and J. This is an important difference from the proposal of section 34.2.3, where we attempted to model the fine-tuning argument with a finitely additive measure that assigns zero probability to each finite interval.

If we can use ν to define our credences, then we have a response to the normalization objection. For any interval I, set

$$Pr(k \in I/H_{Ch}) = \nu(I).$$

Then $Pr(k \in I/H_{Ch}) = Pr(k \in I'/H_{Ch})$ whenever I and I' have equal size, as required. We can extend this idea to define a complete probability model on $\mathbb{R}^+ \times \mathcal{H}$. Just proceed in the same way as in section 34.2.5: first specify families of selection hypotheses, then assign a prior distribution over these families, and finally assign a uniform distribution within each family. If we succeed in defining such a distribution, we have a model for the assumptions of the fine-tuning argument. But that model will be vulnerable to the lottery objection. If we start with unbiased prior probabilities, then the observation of fine-tuning leaves the probability of each unbiased family unchanged. There is no disconfirmation of the chance hypothesis and hence no meaningful confirmation for design.

Our conclusion is that the Bayesian version of the fine-tuning argument faces a serious dilemma. If we reject non-standard measures and impose no cap on the possible value of k, then the fine-tuning argument is defeated by the normalization objection. If we accept non-standard measures or impose a cap, then the fine-tuning argument is invalid because of the lottery objection. An advocate of the fine-tuning argument might try to escape the dilemma by offering independent justification that any rational prior distribution over possible selection hypotheses favours life-permitting values for the fundamental constants. Only in this way can the observation that the constants are right provide meaningful confirmation of design and disconfirmation of chance.

34.3 Pascal's Wager

34.3.1 Introduction

Pascal's Wager ("the Wager") is the argument that, given certain premises, it is rational to "wager for God," which means to take steps to bring about or strengthen one's belief in God. This is a prudential or "pragmatic" argument. It does not appeal to evidence that might

Table 34.1 (Original Wager)

	God exists	*God does not exist*
Wager for God	∞	f_1
Wager against God	f_2	f_3

increase our degree of belief in God. Instead, probabilities share the stage with utilities in a decision-theoretic argument that any rational person should *choose the action* of wagering for God. The main idea is that a decision to take the wager has infinite expected utility, however small the probability (credence) that God exists, so long as that probability is positive and the potential reward is infinite.[28] Hence, wagering for God has expected utility superior to that of all actions that can deliver only (finite) worldly pleasures.

Contemporary expositions of Pascal's Wager use decision matrices such as Table 34.1 (Hájek 2012a). The values f_i represent finite utilities, while the best outcome, salvation, has infinite value. Provided we assign positive probability p to God's existence, we readily see that the expected utility $(p \cdot \infty + (1\text{-}p) \cdot f_1)$ of wagering for God is infinite, while that of wagering against $(p \cdot f_2 + (1\text{-}p) \cdot f_3)$ is finite. If we are rationally required to perform the act that has maximum expected utility, then it follows that we are required to wager for God. Still following Hájek (2012a), the key assumptions are thus:

Premise 1: The decision faced by the agent is appropriately modelled by Table 34.1.
Premise 2: The probability p that *God exists* is a positive (non-infinitesimal) real number: $p > 0$.
Premise 3: Rationality requires choosing the act with maximum expected utility (if there is one).

Hájek (2012a) provides an excellent overview of the major objections to this argument. They call into question each of the three premises, as well as the validity of the argument.

The discussion here will be organized not around the objections *per se* (although they will be prominent), but instead around three themes: *infinite utility*, *infinitesimal probability* and *dynamical approaches* to the Wager. These three themes highlight fruitful areas of intersection between the philosophy of religion and the philosophy of probability. Formal work in all three areas moves Pascal's Wager into new territory and ensures its enduring interest.

34.3.2 Infinite Utility

Pascal's Wager leans heavily on the premise of an infinite reward.[29] But the concept of infinite utility may be incoherent. Indeed, Pascal's followers built decision theory around

[28] Hacking 1972/1994 identifies three distinct arguments in the *Pensées*. Here, I discuss only the third.

[29] Some have given up on infinite utility (Jordan 2006, McClennen 1994, Duff 1986, and Jeffrey 1983) or proposed ways to reformulate the Wager to invoke only finite utilities (Mougin and Sober 1994, Jordan

Table 34.2 (Many gods)

	god$_1$ *exists*	*god*$_2$ *exists*	...	*god*$_n$ *exists*	*No god exists*
Wager for god$_1$	∞	·	...	·	·
Wager for god$_2$	·	∞	...	·	·
...					
Wager for god$_n$	·	·	...	∞	·
Wager against all	·	·	...	·	·

a set of assumptions that "logically excludes infinite utilities" (McClennen 1994).[30] Infinite utility is a double-edged sword: it clears the way for the simple expected utility calculation that constitutes Pascal's main argument, but it also leads to many serious objections.

One of these is the *invalidity* objection: even if the three premises are granted, the argument is invalid if we allow mixed strategies (Jeffrey 1983, Duff 1986, Hájek 2003). Any mixed strategy that assigns positive probability q to *wager-for* and $(1\text{-}q)$ to *wager-against* has infinite expectation. Flipping a coin and wagering only on a result of *Heads* has as great a claim on our rationality as does the pure wager. The conclusion that we ought to wager for God does not follow. Hájek goes further: even a resolute decision to *wager-against* has infinite expectation, provided we make allowance for the possibility of later reversal.

A second is a special case of the notorious *many-gods objection*. In general, this is the objection that the decision matrix should be expanded to include more than one theological possibility.[31] If we are prepared to assign positive probability to one deity, why stop there? Suppose that each of these deities offers an infinite reward to believers. The decision matrix is shown in Table 34.2. The dots represent finite utility values, which are not shown. This is a special case of the many-gods objection because it encompasses only *finitely many* rival deities, and it assumes that all of them are *jealous*: they reward their followers and nobody else. But we are already in trouble because each option (apart from the last) has infinite expectation; hence, there are no grounds for settling upon any wager (even if we rule out mixed strategies).

Schlesinger (in a slightly different context) offers a plausible principle that, if added to decision theory, could help to break the stalemate:

> **Schlesinger's Principle.** In cases where the mathematical expectations are infinite, the criterion for choosing the outcome to bet on is its probability.
>
> (Schlesinger 1994: p. 90).

Applied to Table 34.2, Schlesinger's Principle directs us to wager for the deity to whose existence we assign the highest subjective probability.[32] Schlesinger's Principle also solves

2006, Sobel 1996, Hájek 2003). As the finite versions differ significantly from the one outlined in section 34.3.1 and raise less challenging issues for the philosophy of probability, they are not discussed here.

[30] McClennen 1994 points out that the inclusion of infinite utility violates standard assumptions about preferences needed to justify *Premise 3*, the identification of expected utility maximization as the criterion of rational choice.

[31] Mougin and Sober 1994 present a variant of this objection in which the decision matrix includes alternative background theological commitments, rather than commitments to believe in a deity.

[32] If there is a tie for highest, we are free to choose any in the top group. *Premise 3* in section 34.3.1 should be modified to read: Rationality requires choosing *an* act with maximum expected utility.

the *invalidity* problem: mixed wagers result in lower probability for gaining the infinite prize than the pure wager.[33] The Principle has a further benefit: it allows the Pascalian to wager for a somewhat more humane (less jealous) deity than the one represented in the Original Wager. Suppose that we modify the bottom-left entry in Table 34.1: if God exists and you wager against, there is still a positive probability $q < 1$ for the infinite reward. With this deity, salvation is possible for non-believers but faith is the surest route. Schlesinger's Principle justifies taking the pure wager.

Taken by itself, Schlesinger's Principle looks like an *ad hoc* addition to standard decision theory (Sorensen 1994, Hájek 2003; 2012a). But the Principle is so plausible and helpful for dealing with infinite utility that we should investigate whether there is an extended or modified decision theory in which it holds. We might hope that such a decision theory would provide a satisfactory response to both of the preceding objections.

The problem, then, is how to represent the reward of infinite value. The system that seems best to reflect Pascal's concept of infinity is the *extended real numbers*.[34] This set is formed by adding ∞ and $-\infty$ to the real numbers, extending the ordering so that $-\infty < x < \infty$ for all real x, and extending the arithmetical operations with a few postulates: $x + \infty = \infty$, $x + -\infty = -\infty$, $x \cdot \infty = \infty$ for $x > 0$, and so forth. As Hájek (2003) points out, the postulate $x + \infty = \infty$, which Hájek calls *reflexivity of* ∞ *under addition*, captures Pascal's idea that infinity is absolutely maximal: nothing can be added to it. The postulate $x \cdot \infty = \infty$ for $x > 0$, which Hájek calls *reflexivity of* ∞ *under multiplication*, is used in our initial presentation of the Wager. But as we have just seen, this very postulate makes it impossible to adopt Schlesinger's Principle or to respond effectively to the two objections. If we assume *reflexivity under multiplication*, two outcomes with infinite utility look equally good regardless of any difference in probability.

Hájek (2003) explores a number of sophisticated alternatives for representing infinite utility in the decision matrix. Instead of the extended real numbers, we can make use of a system such as the surreal numbers or the hyperreals $^*\mathbb{R}$, both of which contain infinite numbers. If we represent the value of salvation as an infinite value η and assign probability $p > 0$ to God's existence, then $p \cdot \eta$ is still infinite, which means that the argument of the Wager goes through. Furthermore, $p \cdot \eta < q \cdot \eta$ whenever $p < q$.[35] This means that if expected utility calculations can be computed in the usual way, then Schlesinger's Principle is validated. As Hájek notes, however, within these number systems there is no maximal element, and the postulate of reflexivity of infinity under addition is false. As a result, the choice of a particular infinite number to represent the utility of salvation appears arbitrary: why is your utility η rather than $\eta + 1$? He concludes that these representations are philosophically unsatisfactory.

More recently (2012b), Hájek returns to the extended real numbers—with a twist. He notes that if we replace f_2 with $-\infty$ in Table 1, representing the outcome of wagering against God as infinitely bad if God exists, then we have a response to the invalidity objection. The expectation of *Wager-against* is $-\infty$, the expectation of *Wager-for* is ∞, and the expectation of any mixed strategy is indeterminate. Thus, *Wager-for* is uniquely rational, provided

[33] The same point gives us a possible response to the "delayed belief" objection, on which you decide now to remain agnostic for 50 years and then start believing (Monton 2011).

[34] See Royden 1968 for a brief presentation.

[35] This holds for both the surreal numbers and the hyperreals.

Table 34.3 (Original Wager, Swinburne version)[36]

	God exists	*God does not exist*
Wager for God	1	0
Wager against God	0	0

Table 34.4 (Many gods, Swinburne version)

	god_1 *exists*	god_2 *exists*	...	god_n *exists*	*No god exists*
Wager for god_1	1	0	...	0	0
Wager for god_2	0	1	...	0	0
...					
Wager for god_n	0	0	...	1	0
Wager against all	0	0	...	0	0

that we supplement decision theory with two rules: "prefer anything of expectation ∞ to anything of indeterminate expectation; and prefer anything of indeterminate expectation to anything of expectation $-\infty$" (2012b, p. 181). These rules appear less *ad hoc* than Schlesinger's Principle. While the "negative infinity" strategy does provide a successful response to the invalidity objection, Hájek notes that it appears to be unfaithful to Pascal's theology. There is a further limitation: the use of $-\infty$ provides no help with the many-gods problem, as depicted in Table 34.2. If we introduce $-\infty$ for off-diagonal values, then each option has indeterminate expectation and Hájek's rules provide no guidance. (Of course, it was not Hájek's objective to solve that problem.)

Swinburne (1969) offers an interesting proposal that represents infinite utility as 1 and every finite utility as 0. Following this suggestion, we can represent the decision matrix for the original wager as shown in Table 34.3. Using the same technique, the many-gods problem of Table 34.2 can be represented as shown in Table 34.4. If we make decisions via (naïve) expected utility calculations with these new tables, then we can respond to the two objections raised earlier. With regard to Table 34.3, suppose that p is the probability that God exists. The mixed strategy α(*Wager-for*) + (1-α)(*Wager-against*) has expected utility $p\alpha$, whereas the pure strategy *Wager-for* has expected utility p. This eliminates the invalidity objection. Similarly, for Table 34.4, naïve expected utility calculations lead precisely to the same recommendation as Schlesinger's Principle: decide on the basis of the probabilities. This eliminates the special version of the many-gods problem.

Sobel (1996: p. 19) rightly objects that Table 34.3 (and presumably Table 34.4) misrepresents the relations among finite outcomes. Different finite utilities all collapse to 0.

[36] Swinburne puts -1 for the value $-\infty$ and considers a three-column decision table, but I set aside these complications.

This criticism is valid if our intention is to represent an agent's preference ranking with a one-dimensional utility function. Of course, the same criticism applies to Table 34.1: it does a fine job of representing an agent's preferences among finite outcomes, but it fails to discriminate among outcomes with infinite utility which are all represented as having the same value, ∞.

There is a way to think about Swinburne's proposal that evades Sobel's criticism. This is the *relative-utilities* approach developed in Bartha (2007). Rather than attempting to characterize absolutely infinite utility, the idea is to define *infinite relative utility*. Start with the revealed preferences of a Pascalian agent who prefers salvation infinitely over finite existence. What this means is that for the Pascalian, life as a non-believer (B) is inferior to *any* gamble between salvation (A) and life as a deluded believer (Z), so long as there is some positive real-valued probability, however slight, of obtaining A. In general, an agent's utility for A is infinite relative to B with "base-point" (worst-case outcome) Z just in case the agent prefers *any* non-trivial gamble between A and Z to B. That is:

Infinite relative utility.

$$U(A, B; Z) = \infty \quad \text{if and only if } B \preccurlyeq [pA, (1\text{-}p)Z] \text{ for all real } 0 < p \leq 1.$$

Here, $\preccurlyeq$ denotes a weak ordering on preferences, while $[pA, (1\text{-}p)Z]$ denotes a gamble with probability p of outcome A and $(1\text{-}p)$ of outcome Z. Notice that infinite relative utility is defined in terms of preferences among ordinary, well-defined gambles. Relative utilities can also be *finite*: if the agent is indifferent between B and a particular gamble $[\alpha A, (1-\alpha)Z]$, then the relative utility of B is α (with respect to gambles between A and Z). We always pick the base-point Z to be no better than A or B, so that relative utilities are non-negative extended real numbers between 0 and ∞ (inclusive).

On the relative-utilities approach, Tables 34.3 and 34.4 are tables of relative utilities: they represent utility values *relative to the best outcome* in the table. (The base-point can be any worldly outcome inferior to those in the table.) In just the same way, Tables 34.1 and 34.2 represent utility values *relative to some finite outcome*. All of these tables convey only *partial information* about the agent's preferences. Table 34.1 provides the view from the ground floor. It indicates utility ratios relative to ordinary finite outcomes, but is unable to discriminate between the far-off infinite outcomes, which are all represented as ∞. Table 34.3 provides the view from above. It indicates utility ratios relative to the best outcome, but is unable to discriminate among the finite outcomes, which all come out as 0. The response to Sobel's objection is that we can use *either* representation, as appropriate, to make decisions. We don't regard either table as a complete representation of the agent's preferences; that information is contained in the three-place function U. So it is legitimate to use Tables 34.3 and 34.4 to respond to the invalidity objection and to solve the special case of the many-gods problem.

As a way to represent infinite utility, relative utilities and relative decision matrices offer some additional advantages over the number systems considered (and rejected) by Hájek. First, since they only make use of the extended real numbers, they are compatible with Pascal's intuition that the utility of salvation is absolutely maximal. Secondly, the injunction to maximize expected relative utility is supported by a representation theorem. Thirdly, they

are easy to use: form a decision table by putting 1 for infinite outcomes[37] and 0 for finite outcomes, compute (relative) expected utilities in the usual way, and select a strategy that maximizes (relative) expected utility.

A major limitation, however, is that relative utilities are defined in terms of standard real-valued probabilities. No infinitesimals are allowed! This brings us to our second theme.

34.3.3 Infinitesimal Probability

Pascal's argument relies upon the fact that $p \cdot \infty = \infty$ for any positive real number p. But it does not seem unreasonable to refuse to grant that the probability of God's existence exceeds p for any positive real value p, yet (perhaps with a nod to *Regularity*) to leave the door slightly open by assigning $Pr(God\ exists) = \varepsilon$ for a positive *infinitesimal* value ε. What becomes of the *Wager* if we take this path?

As Hájek (2012a) notes, naïve expected utility reasoning gets us nowhere. For the first row of Table 34.1, we get:

$$\text{EU(wager for God)} = \varepsilon \cdot \infty + (1 - \varepsilon) f_1.$$

What is the product $\varepsilon \cdot \infty$? If ε belongs to ${}^*\mathbb{R}$ (the hyperreals) and ∞ belongs to the extended real numbers, then the multiplication is simply undefined. If we replace ∞ with an infinite hyperreal η, then the answer is indeterminate: $\varepsilon \cdot \eta$ could turn out to be infinitesimal, a finite real number, or infinite.[38] In section 34.3.2, we saw that it is unclear which infinite hyperreal η should represent the value of the infinite reward; here, it is unclear which infinitesimal ε best represents the probability of God's existence. Until we resolve both points, we are not in a position to make a comparison between the expected utilities of wagering for and against God.

The situation vis-à-vis infinitesimal probability is different in decision theory from what it is in a purely epistemic setting. In the application of non-standard measure to the fine-tuning argument (section 34.2.6), the choice of infinitesimal is unimportant. Any infinitesimal serves equally well as a way to generate a uniform distribution over the real numbers. In a decision-theoretic setting with infinite utilities, the value of the infinitesimal matters. Traditionally, there is a link between credences and betting behaviour (Ramsey 1931, Jeffrey 1983). If we allow infinitesimal probabilities for God's existence, then decision theory should tell us which bets involving God's existence are rational, even when infinite utilities are in play. But no decision theory does this.[39]

Oppy (1991) argues that rationality *requires* that the probability assigned to God's existence be infinitesimal, independently of any consideration of utilities. The conclusion

[37] There is an argument for representing all infinite pure outcomes as 1, but this is not mandatory; see Bartha 2007; 2012.

[38] For example, we could have $\varepsilon = \eta^{-2}$, $\varepsilon = \eta^{-1}$ or $\varepsilon \cdot \eta^{-1/2}$. Here, we consider R as a subfield of *R.

[39] Hammond 1999 has a rigorous decision theory using the smallest possible set of infinitesimal probabilities, namely, those in the non-Archimedean field R(ε). But Hammond's decision theory is restricted to finite utility values. No rigorous decision theory combines non-Archimedean utility and non-Archimedean probability.

derives from an infinite version of the many-gods problem.[40] Oppy asks us to consider, for each positive integer *n*, the hypothesis that there are "*n* deities (all much like the Christian God) who reward all and only those people who believe that there are *n* deities who are much like the Christian God." Symmetry, together with an application of the *Principle of Indifference*, requires an assignment of equal probability to each theological possibility. If these probabilities are to be non-zero, infinitesimals appear inevitable.

Critics of this reasoning, such as Jordan (2006), argue that these gerrymandered deities ("philosophers' fictions" is Jordan's term) have no plausibility and may be dismissed. But suppose we concede Oppy's point: the assignment of infinitesimal probability ε is mandatory. In this case, we cannot compute expected utilities and the reasoning of Pascal's Wager collapses.

The best response for the Pascalian, perhaps, is to offer a different interpretation of the Wager than the presentation in section 34.3.1. On the standard presentation, the starting points are that salvation has infinite utility and that God's existence has positive probability; the conclusion is that it is rational to wager for God. As we have just seen, we won't get anywhere with these premises if the probability is infinitesimal. The alternative interpretation is to view the problem as a puzzle about representation, rather than rationality. The starting point is the Pascalian agent's preference for wagering for God over any worldly prize, a preference that persists even if the probability of God's existence is infinitesimal. The challenge is to find a representation for these preferences that shows that they are rationally permissible. In other words, the problem is to formulate a decision theory within which Pascalian choices satisfy a criterion of expected utility maximization. This decision theory need only accommodate *some* infinite utilities and *some* infinitesimal probabilities without generating indeterminate expectations.

34.3.4 Dynamical Approaches to the Wager

Pascal's Wager is presented as a piece of practical reasoning to be performed once. If we accept the premises and follow the reasoning, then we should act by taking measures to bring about belief. Yet there are reasons to adopt a dynamical view of the Wager. Rather than regarding the Wager as an argument about a single choice, we should treat the probabilities and decisions in a Pascalian decision problem as evolving. In this section, I briefly review two dynamical approaches to Pascal's Wager.

The first comes in a critique by Easwaran and Monton (2012) of Duff and Hájek's "invalidity" objection (see section 34.3.2). The invalidity objection, once again, is that the criterion of utility maximization does not single out the pure wager because every mixed strategy also has infinite expectation. In their critique, Easwaran and Monton concede that the mixed strategies have infinite expected utility. Their starting point is that it is at least rational to choose *some* mixed strategy (rather than *Wager-against*) at each moment, so long as one has not yet taken the wager. They show that a *sequence* of such choices perhaps implemented using some complex set of coin tosses, results in taking the wager

[40] Gale 1991 and Jordan 2006 present numerous similar examples.

with probability 1.[41] A rational supertasker capable of performing all of these tosses will appreciate the inevitability of wagering, skip the coin tosses and just wager for God at the outset. Easwaran and Monton then argue that a rational agent is bound to do what is rational for her ideal (supertasker) counterpart. Furthermore, for those who take the wager, "there is no going back"; otherwise, Pascal's argument is pointless. They conclude that a rational agent should take the pure wager, rather than pursuing a mixed strategy or wagering against God.

There is a minor flaw in this reasoning. The initial assumption that it is rational to pursue *some* mixed strategy ignores Hájek's observation that *every available action*, even a determined resolution to wager against God, has infinite expectation, so long as there is some probability that one will end up wagering for God. If we counter that every available action amounts to a mixed strategy, we are no longer discussing what it is rational to do. Instead, we merely have a prediction that the supertasker *will* eventually take the wager. Without the link to rationality, the assumption that what the supertasker does is binding on the actual agent does not follow. To avoid the flaw, we must assume that eschewing the mixed strategies ensures that one does not take the wager (and hence gains only finite expected utility).

Setting aside this small concern, Easwaran and Monton's argument is distinctive because it implicitly defines a dynamics for Pascalian updating. Agents have two possible states at any moment: *Accept* or *Don't accept* the wager. There is no possibility of change from *Accept* at moment t to *Don't accept* at a later moment t'. At every moment, however, there is pressure to change from *Don't accept* to *Accept*. On this model, there is an equilibrium state: *Accept*. Further, it is a stable equilibrium. The proof of convergence to this equilibrium is straightforward but important: without the assumption of uncountably many decisions, we don't have convergence with probability 1.

A different kind of Pascalian dynamics is reflected in the "many wagers" model developed by Bartha (2012). The model is motivated by another form of the many-gods objection. Philosophers have pointed out that Pascal's argument can fail if we assign positive probability to gods who reward people other than their followers, or to gods who do not reward their followers at all.[42] If we change our probabilities enough, atheism can become the most prudent wager even if God's existence has positive probability. This could be because we assign a very high subjective probability to a god who rewards atheists or for some other reason (Mougin and Sober 1994).

To illustrate the problem, consider the following example, which pits Pascal's jealous god against a grouchy god who rewards only atheists. The situation is illustrated in Table 34.5. Assume positive probabilities p_J, p_G and p_A for each column. The pure strategies "Wager for jealous" and "Wager against all" (and, of course, mixtures of these two strategies) have equal infinite expectation. There is no clear choice.

We can make some headway by using the relative decision matrices of section 34.3.2, or by using any decision theory that validates Schlesinger's Principle. There is a clear choice: *Wager for jealous god* if $p_J > p_G$ and *Wager against all* if $p_G > p_J$. There is, however, something puzzling about the latter result. The basis for your decision to wager against all is that $p_G > p_J$. You take the grouchy god more seriously than the jealous god. Yet to wager against all

[41] The proof assumes that choices are made (or at least are available) at an *uncountable* set of moments, such as an interval of real numbers.

[42] For a lengthy list of criticisms of this sort, see Jordan 2006 p. 75n.

Table 34.5 (Jealous vs. Grouchy)

	Jealous god exists	Grouchy god exists	No god exists
Wager for jealous god	∞	·	·
Wager for grouchy god	·	·	·
Wager against all	·	∞	·

means to take steps that will lower your credence in the grouchy god—albeit in the jealous god as well. So long as the inequality $p_G > p_J$ is maintained, you persist in wagering against all. Still, persistence leads to a situation where $p_J = p_G = 0$ and $p_A = 1$, undermining your original reason for wagering against all. Furthermore, the details of the convergence are important, since if the inequality $p_G > p_J$ ceases to hold at any point, you should change your wager.

In short, "wagering against all" may be self-undermining in a way that wagering for the jealous god is not. Furthermore, as in Easwaran and Monton's reasoning, you can foresee where things are headed when you begin your deliberations. You recognize the tension between the assumptions that support your initial reasoning and your anticipated final epistemic state. This tension should motivate you to look not just at the expected payoffs for a single decision, but also at the sustainability of your probability assignments (and your decisions based on those assignments).

On the "many wagers" model, you assign each pure wager a weight that matches your current subjective probability for the corresponding deity (or against all deities). Starting with a set of initial probabilities, you compute relative expected utilities for the various wagers. Your probability increases for deities whose associated wager has higher than average relative expected utility, and decreases for wagers that do worse than average. You then repeat the process with the modified probabilities. The updating rule is modelled on the Replicator Dynamics of evolutionary game theory (Skyrms 1996). Based on relative expectations, the probabilities rise or fall until they approach equilibrium: a set of probability assignments that is invariant under the updating rule.

By way of illustration, consider first the original wager, as represented by Table 34.3. The relative expectation for rejecting the wager is 0. So long as your initial probability p for God's existence is positive, this probability rises to 1 after a single iteration. The only equilibria are $p = 0$ or $p = 1$, and only the latter is a stable equilibrium (i.e., if small changes are made to these probability values, the original values are restored by the updating rule). If the final distribution must be a stable equilibrium, then $p = 1$ is the only possibility; it is the only possible epistemic limit for someone who begins with $p > 0$.

As a further example, take the case of *Jealous vs. Grouchy*. Obtain the relative-utilities version of Table 34.5 by replacing ∞ with 1 and · with 0. Compute relative expectations, starting with any initial set of positive probabilities p_J, p_G, and p_A. We find that p_G drops to 0 after one application of the updating rule, because the expectation of wagering for a grouchy god is 0. After two applications, p_A drops to 0 as well. The only equilibrium values for (p_J, p_G, p_A) are (1, 0, 0), (0, 0, 1) and (0, 1, 0), and only the first of these is a stable equilibrium. If you have any inclination to believe in the jealous god (i.e., $p_J > 0$ initially),

the model leads to full belief. This is a much stronger result than in the earlier discussion of this example, where belief in the jealous god is vindicated only if $p_J > p_G$ initially.

The most important idea here, as in the Easwaran/Monton model, is that the beliefs and wagering decisions in a Pascalian problem evolve in response to pressure. The concept of equilibrium offers a fruitful criterion for evaluating and critiquing the subjective probability assignments either in the original argument or in complex many-gods scenarios.

34.4 Conclusion

Our understanding of arguments in the philosophy of religion has been, and continues to be, deepened by extensions of Bayesian ideas and by the use of non-Bayesian models; conversely, the philosophy of religion provides important motivation and examples that help us to refine ideas in formal epistemology. Both points are evident from our discussion of the fine-tuning argument and Pascal's Wager.

In the case of the fine-tuning argument, the finite models of section 34.2.5 and their extension (via non-standard measure) to the infinite case provide both a possible response to the normalization objection and a challenge in the form of the lottery objection. In the reverse direction, as the discussion of observation-selection effects in section 34.2.4 shows, careful scrutiny of the fine-tuning argument yields insight into important differences between likelihoodist and Bayesian approaches to evidence.

As for Pascal's Wager, section 34.3.2 shows that a variety of formal devices for representing infinite utility offer a means of responding to the venerable many-gods objection, as well as to the invalidity objection posed by Duff and Hájek. Standard decision theory, as we saw, does not provide adequate resources to meet either challenge. Conversely, decision theorists see Pascal's Wager as a gateway to infinite decision theory, the full development of which will require concepts of infinite value, infinitesimal probability, and possibly some type of evolutionary dynamics.

Sprouting from a mixture of analogies, vague intuitions or principles, and mathematical puzzles, formal models thus help us both to sharpen old objections and to open new perspectives. The same point applies to other classic arguments in the philosophy of religion, such as Hume's famous critique of miraculous testimony. We have every reason to expect an enduring vigorous relationship between philosophy of religion and philosophy of probability

Acknowledgments

For valuable comments and discussion, I thank Chris Stephens, Richard Johns, Stefan Lukits, Adam Morton, Elliott Sober, Alexander Pruss, Chris Hitchcock, and Alan Hájek.

APPENDIX

Proof of (18) (section 34.2.5)

Proposition: Given the distribution *Pr*, sets D_n and evidence *O* as defined in section 34.2.5, $Pr(D_n/O) = Pr(D_n)$ for each *n*.

Proof: If *J* has *n* members, then $Pr(O/J) = 1/n$ if $k^* \in J$, and 0 otherwise.

Within D_n, there are $\binom{N-1}{n-1}$ sets *J* such that $k^* \in J$. It follows that

$$\begin{aligned} Pr(O/D_n) &= \sum_{J \in D_n} \Pr(O/J) \Pr(J/D_n) \\ &= \binom{N-1}{n-1} \left(\frac{1}{n}\right) \frac{1}{\binom{N}{n}} \\ &= 1/N \end{aligned}$$

(as we might expect, given the uniform distribution within D_n). Hence also,

$$Pr(O) = \Sigma Pr(O/D_n)\, Pr(D_n) = (\Sigma \alpha_n)/N = 1/N.$$

By Bayes' Theorem, it follows trivially that

$$Pr(D_n/O) = \frac{\Pr(O/D_n)\Pr(D_n)}{\Pr(O)} = Pr(D_n).$$

References

Barrow, J. D. (2001) Cosmology, Life, and the Anthropic Principle. *Annals of the New York Academy of Sciences*. 950. 1. pp. 139–53.

Barrow, J. D. and Tipler, F. J. (1986) *The Anthropic Cosmological Principle*. Oxford: Oxford University Press.

Bartha, P. (2007) Taking Stock of Infinite Value: Pascal's Wager and Relative Utilities. *Synthese*. 154. pp. 5–52.

Bartha, P. (2012) Many Gods, Many Wagers: Pascal's Wager Meets the Replicator Dynamics. In Chandler, J. and Harrison, V. (eds.) *Probability in the Philosophy of Religion*. pp. 187–206. Oxford: Oxford University Press.

Bayes, T. (1763) An Essay towards Solving a Problem in the Doctrine of Chances. *Philosophical Transactions of the Royal Society (London)*. 53. pp. 370–418.

Bostrom, N. (2002) *Anthropic Bias: Observation Selection Effects in Science and Philosophy*. New York, NY and London: Routledge.

Carter, B. (1974/1989) Large Number Coincidences and the Anthropic Principle in Cosmology. IAU Symposium 63: Confrontation of Cosmological Theories with Observational

Data. pp. 291–298. Dordrecht: Reidel. (Reprinted in Leslie, J. (ed.), *Universes*. London: Routledge.)
Collins, R. (2009) The Teleological Argument: An Exploration of the Fine-Tuning of the Universe. In Craig, W. L. and Moreland, J. P. (eds.) *The Blackwell Companion to Natural Theology*. pp. 202–81. Oxford: Blackwell.
Colyvan, M., Garfield, J., and Priest, G. (2005) Problems with the Argument from Fine-tuning. *Synthese*. 145. pp. 325–38.
Craig, W. L. (2003) Design and the Anthropic Fine-tuning of the Universe. In Manson, N. (ed.) *God and Design*. pp. 155–77. New York, NY and London: Routledge.
Crupi, V. and Tentori, K. (2016) Confirmation Theory. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Dembski, W. (1998) *The Design Inference: Eliminating Chance through Small Probabilities*. Cambridge: Cambridge University Press.
Dembski, W. (2003) The Chance of the Gaps. In Manson, N. (ed.) *God and Design*. pp. 251–74. New York, NY and London: Routledge,
Duff, A. (1986) Pascal's Wager and Infinite Utilities. *Analysis*. 46. pp. 107–9.
Easwaran, K. and Monton, B. (2012) Mixed Strategies, Uncountable Times, and Pascal's Wager: A Reply to Robertson. *Analysis*. 72. 4. pp. 681–5.
Eddington, A. (1939) *The Philosophy of Physical Science*. Cambridge: Cambridge University Press.
Edwards, A. (1972) *Likelihood*. Cambridge: Cambridge University Press.
Gale, R. (1991) *On the Nature and Existence of God*. Cambridge: Cambridge University Press.
Gillies, D. (2000) *Philosophical Theories of Probability*. London: Routledge.
Goldblatt, R. (1998) *Lectures on the Hyperreals*. New York, NY: Springer-Verlag.
Hacking, I. (1965) *The Logic of Statistical Inference*. Cambridge: Cambridge University Press.
Hacking, I. (1972/1994) The Logic of Pascal's Wager. *American Philosophical Quarterly*. 9. 2. pp. 186–92. (Reprinted in Jordan, J. (ed.) *Gambling on God: Essays on Pascal's Wager*. pp. 21–9. Lanham, MD: Rowman & Littlefield.)
Hájek, A. (2003) Waging War on Pascal's Wager. *Philosophical Review*. 112. 1. pp. 27–56.
Hájek, A. (2011) Interpretations of Probability. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/probability-interpret/. [Accessed 25 Sep 2015.]
Hájek, A. (2012a) Pascal's Wager. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/pascal-wager/. [Accessed 25 Sep 2015.]
Hájek, A. (2012b) Blaise and Bayes. In Chandler, J. and Harrison, V. (eds.) *Probability in the Philosophy of Religion*. pp. 167–86. Oxford: Oxford University Press.
Hájek, A. (2013) Staying Regular? draft manuscript. [Online] Available from: http://philosophy.anu.edu.au/sites/default/files/Staying%20Regular.December%2028.2012.pdf. [Accessed 25 Sep 2015.]
Halpern, J. (2010) Lexicographic Probability, Conditional Probability, and Nonstandard Probability. *Games and Economic Behavior*. 68. 1. pp. 155–79.
Hammond, P. (1999) Non-Archimedean Subjective Probabilities in Decision Theory and Games. *Mathematical Social Sciences*. 38. 2. pp. 139–56.
Howson, C. and Urbach, P. (1993) *Scientific Reasoning: The Bayesian Approach*. 2nd edition. La Salle, IL: Open Court Press.

Hurd, A. E. and Loeb, P. (1985) *An Introduction to Nonstandard Real Analysis*. Orlando, FL: Academic Press.
Jeffrey, R. C. (1983) *The Logic of Decision*. 2nd edition. Chicago, IL: University of Chicago Press.
Jordan, J., (ed.) (1994). *Gambling on God: Essays on Pascal's Wager*. Lanham, MD: Rowman & Littlefield.
Jordan, J. (2006) *Pascal's Wager: Pragmatic Arguments and Belief in God*. Oxford: Clarendon Press.
Kadane, J., Schervish, M., and Seidenfeld, T. (1996) Reasoning to a Foregone Conclusion. *Journal of the American Statistical Association*. 91. 435. pp. 1228–35.
Kotzen, M. (2016) Probability in Epistemology. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Laplace, P. S. (1820) *Oeuvres complètes*. Paris: M. V. Courcier.
Leslie, J. (1989) *Universes*. London: Routledge.
Leslie, J. (ed.) (1998) *Modern Cosmology and Philosophy*. New York, NY: Prometheus Books.
McClennen, E. (1994) Pascal's Wager and Finite Decision Theory. In Jordan, J. (ed.) *Gambling on God: Essays on Pascal's Wager*. pp. 115–37. Lanham, MD: Rowman & Littlefield.
McGrew, T. (2011) Miracles. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/miracles/. [Accessed 25 Sep 2015.]
McGrew, T., McGrew, L., and Vestrup, E. (2001) Probabilities and the Fine-Tuning Argument: A Sceptical View. *Mind*. 110. pp. 1027–37.
McMullin, E. (1993) Indifference Principle and Anthropic Principle in Cosmology. *Studies in the History and Philosophy of Science*. 24. pp. 359–89.
Monton, B. (2011) Mixed Strategies Can't Evade Pascal's Wager. *Analysis*. 71. 4. pp. 642–5.
Mougin, G. and Sober, E. (1994) Betting Against Pascal's Wager. *Noûs*. 28. pp. 382–95.
Oppy, G. (1991) On Rescher on Pascal's Wager. *International Journal for Philosophy of Religion*. 30. 3. pp. 159–68.
Paley, W. (1802) *Natural Theology, or, Evidences of the Existence and Attributes of the Deity, Collected from the Appearances of Nature*. London: Rivington.
Pascal, B. (1670/1910) *Pensées*. Translated from the French by W. F. Trotter. London: Dent.
Plantinga, A. (2011) *Where the Conflict Really Lies: Science, Religion and Naturalism*. New York, NY: Oxford University Press.
Price, R. (1767/1768) *Four Dissertations*. 2nd edition. London: A. Millar and T. Cadell.
Ramsey, Frank (1931) Truth and Probability. In Braithwaite, R. B. (ed.) *The Foundations of Mathematics and other Logical Essays*. pp. 156–98. London: Routledge and Kegan Paul.
Ross, H. (1998) Big Bang Model Refined by Fire. In Dembski, W. (ed.) *Mere Creation*. pp. 363–84. Downer's Grove, IL: InterVarsity Press.
Royall, R. (1997) *Statistical Evidence: A Likelihood Paradigm*. London: Chapman and Hall.
Royall, R. (2004) The Likelihood Paradigm for Statistical Evidence. In Taper, M. L. and Lele, S. R. (eds.) *The Nature of Scientific Evidence*. pp. 119–52. Chicago, IL: University of Chicago Press.
Royden, H. L. (1968) *Real Analysis*. 2nd edition. New York, NY: Macmillan.
Russell, P. (2013) Hume on Religion. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/hume-religion/. [Accessed 25 Sep 2015.]
Schlesinger, G. (1994) A Central Theistic Argument. In Jordan, J. (ed.) *Gambling on God: Essays on Pascal's Wager*. pp. 83–99. Lanham, MD: Rowman & Littlefield.

Skyrms, B. (1996) *Evolution of the Social Contract*. Cambridge: Cambridge University Press.
Smolin, L. (1997) *The Life of the Cosmos*. Oxford: Oxford University Press.
Sobel, H. (1996) Pascalian Wagers. *Synthese*. 108. pp. 11–61.
Sober, E. (2002) Bayesianism: Its Scope and Limits. In Swinburne, R. (ed.) *Bayes' Theorem, Proceedings of the British Academy Press*. Vol. 113. pp. 21–38.
Sober, E. (2004) The Design Argument. In Mann, W. (ed.) *The Blackwell Guide to the Philosophy of Religion*. pp. 117–47. Oxford: Blackwell.
Sober, E. (2009) Absence of Evidence and Evidence of Absence: Evidential Transitivity in Connection with Fossils, Fishing, Fine-Tuning, and Firing Squads. *Philosophical Studies*. 143. 1. pp. 63–90.
Sorensen, R. (1994) Infinite Decision Theory. In Jordan, J. (ed.) *Gambling on God: Essays on Pascal's Wager*. pp. 139–59. Lanham, MD: Rowman & Littlefield.
Sprenger, J. (2016) Bayesianism vs. Frequentism in Statistical Inference. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.
Swinburne, R. (1969) The Christian Wager. *Religious Studies*. 4. 2. pp. 217–28.
Swinburne, R. (1990) Argument from the Fine-tuning of the Universe. In Leslie, J. (ed.) *Physical Cosmology and Philosophy*. pp. 160–79. New York, NY: Macmillan.
Swinburne, R. (2004) *The Existence of God*. 2nd edition. Oxford: Clarendon Press.
van Fraassen, B. (1989) *Laws and Symmetry*. Oxford: Clarendon Press.
Weisberg, J. (2005) Firing Squads and Fine-Tuning: Sober on the Design Argument. *British Journal for the Philosophy of Science*. 56. pp. 809–21.
Zynda, L. (2016) Subjectivism. In Hájek, A. and Hitchcock, C. (eds.) *Oxford Handbook of Probability and Philosophy*. Oxford: Oxford University Press.

CHAPTER 35

PROBABILITY IN PHILOSOPHY OF LANGUAGE

ERIC SWANSON

Philosophers of language investigate foundational questions about the nature of linguistic communication. Much language that uses probabilistic vocabulary—in particular, what I'll call *the language of subjective uncertainty*—bears on these questions in important and distinctive ways. Understanding better how the language of subjective uncertainty works will help us get a better picture of the general features of communication and of the vehicles of communication.

Early analytic philosophers of language such as Frege, Russell, and the Wittgenstein of the *Tractatus* were principally interested in mathematical language, where there is little need for probabilistic expressions. But in the middle of the twentieth century many philosophers of language began to pay attention to a wider variety of language uses. J. L. Austin, for example, opines that "It was for too long the assumption of philosophers that the business of a 'statement' can only be to 'describe' some state of affairs, or to 'state some fact', which it must do either truly or falsely" (1962: p. 1). On many recent views, when we communicate from a position of significant subjective uncertainty, we sometimes use language in a way that doesn't describe how the world is. The language of subjective uncertainty is, on such views, like the later Wittgenstein's primitive "slab language" at least in that both are used to communicate without being used to describe.

Here is an example. Suppose Alice is curious about whether Bob won a deterministic raffle. I know that the raffle was fair, and that Bob purchased four of the hundred tickets that were sold. I might say

(1) There's a 4% chance that Bob won the raffle.

I would thereby be communicating from a position of significant subjective uncertainty, and I (arguably) would not be trying to describe how the world really is. After all, I know that either (2) or (3) is the correct relevant *description* of the world:

(2) Bob won the raffle.
(3) Bob didn't win the raffle.

But because I don't know whether (2) or (3) is true, I fall back on the probabilistic, hedged language of (1). I am not trying to communicate the objective chance of Bob winning the raffle, since the raffle was deterministic, making the objective chance that Bob won either 0 or 1. We might instead say that I communicate something like an *estimate* of the truth value of (2).[1] But I am (arguably) not trying to describe or represent the world because there is (arguably) no way the world could be that would make the estimate that I communicate with (1) true or false.

What's at stake here? Prima facie, when we make sincere assertions we aim to represent how things are. But perhaps this isn't quite right, if the language of subjective uncertainty isn't in the business of representing the world. Similarly, many have thought it plausible that an important vehicle of communication—the "content" of a sincere assertion—simply distinguishes between ways the world might be, or between ways the world might be thought to be, or between ways the world might be represented as being. But perhaps this isn't quite right, if these thoughts don't generalize to the language of subjective uncertainty. Disregarding the language of subjective uncertainty may encourage a distorted picture of the vehicles of communication.

Anticipating this dialectic, many theorists have claimed that the language of subjective uncertainty merely indicates something about the speaker's attitudes, signaling perhaps the "reservations to which [some statement] is subject" (Austin 1962: p. 3). This kind of view often treats the language of subjective uncertainty as a sort of side comment that is semantically and syntactically isolated from the language on which it comments. But recent work has shown that the language of subjective uncertainty enters into extensive semantic and syntactic relations with other language. So the language of subjective uncertainty must be analyzed by any complete compositional (or largely compositional) semantic theory, and such a theory must account for the ways in which it interacts with other language. Because traditional compositional semantic theories traffic in propositions (as opposed to, say, probability spaces), this makes it more challenging to explain how and in what sense such language could be used to communicate subjective uncertainty without thereby describing the world. On the one hand, we want compositionality; on the other hand, we want vehicles of communication that successfully model the states associated with the language of subjective uncertainty. But it has long been unclear whether these two desires can be jointly satisfied.

In this chapter I take conditionals as a case study, to illustrate how philosophical thought on the language of subjective uncertainty has evolved. As I use the term, the language of subjective uncertainty includes conditionals, "unless" claims, epistemic modals, epistemic comparatives, and so on, since speakers often use such expressions when communicating from positions of relevant subjective uncertainty. (By contrast, the use of probabilistic vocabulary to convey information about objective chance or as a shorthand representation of relative frequencies is not of special interest to philosophers of language, and I leave such uses aside here.) Conditionals are the most instructive case to study because of the extensive literature discussing them. That literature has long taken into account both arguments that

[1] On estimates of truth value, see Jeffrey 1986 and Joyce 1998; 2005.

the content or semantic value[2] of conditionals must be non-propositional, and arguments that conditionals interact with other linguistic items in systematic ways. At the end of the chapter I return to the language of subjective uncertainty more broadly construed.

35.1 Some Background on Conditionals

While it is difficult to draw precise distinctions here, it is helpful to call conditionals such as (4) *indicative*, and conditionals like (5) *subjunctive* or *counterfactual* (Adams 1970: p. 90).

(4) If Oswald did not kill Kennedy, then someone else did.
(5) If Oswald had not killed Kennedy, then someone else would have.

Clearly there are important semantic differences between (4) and (5). Most people familiar with the Kennedy assassination believe (4). By contrast, only those who believe there were backup assassins believe (5). So there is some important distinction between "'indicative" conditionals such as (4) and "subjunctive" conditionals such as (5), even though many conditionals are difficult to classify as falling on one side or the other.

Conditionals are interesting for a host of reasons. They help us communicate not about how things actually are, but about how things are (or would be) *under some supposition*. And communicating about how things are (or would be) under a particular supposition is incredibly useful. It is difficult to do much deciding, planning, and coordinating without using conditionals; similarly it is difficult to do much work in philosophy without using conditionals or closely related constructions. Most importantly for present purposes, however, probability—in particular conditional probability—seems intimately connected to indicative conditionals. Frank Ramsey articulated this thought in a way (now known as "the Ramsey test") that connects it back to supposition:

> If two people are arguing 'If p will q?' and are both in doubt as to p, they are adding p hypothetically to their stock of knowledge and arguing on that basis about q; . . . . We can say they are fixing their degrees of belief in q given p
>
> (1931: p. 155).

Ramsey's phrase "degrees of belief in q given p" is somewhat ambiguous. That said, Ernest Adams articulates one extremely influential reading:

> *. . . the probability of an indicative conditional of the form "if A is the case then B is" is a conditional probability.* [In other words] . . . the probability of "if A then B" should equal the *ratio* of the probability of "A and B" to the probability of A . . .
>
> (1975: p. 3).

[2] The literature isn't as clear as it should be on the distinction between content and semantic value, although I don't have space to take up the issue here. See Yalcin 2014 for a helpful discussion of this distinction.

While the idea of connecting indicative conditionals to conditional probabilities is prima facie very attractive, it has proven surprisingly difficult to spell out the details.

Sections 35.2 through 35.5 consider four families of analyses of indicative conditionals like (4). According to the first two families of analyses, these conditionals have truth conditions. But only on the first are they truth-functional—that is, only on the first are their truth values wholly determined by the truth values of their antecedents and consequents. Analyses in the second family try to connect indicative conditionals to conditional probability, giving them non-truth-functional truth conditions. The third and fourth families of analyses deny that indicative conditionals have truth conditions, but differ in the ways in which they account for apparent semantic compositionality. All the analyses I consider aim to capture the apparent connection between natural language conditionals and conditional probability. The different ways in which the analyses try to do this constitute different ways of thinking about the general features of communication and about the vehicles of communication.

35.2 Truth-functional Theories

It is generally agreed that if the natural language indicative "if" is a truth-functional connective, then it is the material conditional—true if the antecedent is false, the consequent is true, or both; false otherwise.[3] A naïve material conditional analysis predicts the truth of such odd conditionals as

(6) If the moon is made of cheese, then I had oatmeal for breakfast.

But since some true sentences aren't appropriately assertible, (6) might sound strange because it is misleading, or conversationally infelicitous in some other way.

The canonical starting point for approaches in this family is Paul Grice's argument that when a speaker uses the indicative conditional she conversationally implicates "that there are non-truth-functional grounds for accepting $p \supset q$" (1987: p. 58, originally delivered in 1967; see also Thomson 1990). Those grounds might be something like "p would, in the circumstances, be a good reason for q" or "q is inferable from p". (6) is odd because there aren't such grounds for it.

On David Lewis's elaboration of Grice's theory, the assertibility of "If p, q" comes in degrees. That assertibility is equal to the subjective probability of $p \supset q$ minus a discounting factor, which is the product of two other factors: the probability of the conditional's vacuity (owing to the falsity of the antecedent) and the probability of the consequent's falsity conditional on the antecedent's truth (Lewis 1976, pp. 142–3; cf. Lewis 1973, p. 72). In other words, Lewis holds that the assertibility of "If p, q" is equal to

$$P(p \supset q) - P(\neg p) \cdot \frac{P(\neg q \wedge p)}{P(p)},$$

[3] See, e.g., Edgington 1995: pp. 242–3. Gibbard 1981: pp. 234–5 argues that if indicative conditionals have truth conditions then those truth conditions must be those of the material conditional; see Kratzer 1986 and Gillies 2009 for responses.

which just is the conditional probability of q on p. This connection between assertibility and conditional probability is very attractive.[4]

Frank Jackson criticizes Grice's approach and Lewis's elaboration of it on several grounds. Most importantly for our purposes, he argues that it may be better to assert the *weaker* claim W even if W is equiprobable to the stronger (and no more prolix) claim S. In particular, if W is more likely than S to prove useful as new information comes in, then there might be greater expected value in asserting W rather than S. In his 1979 Jackson calls P *robust* with respect to I iff the probability of P and the conditional probability of P on I are "close and high" (p. 6), and argues that while "High probability is an important ingredient in assertibility … so is robustness" (p. 12). Jackson argues that the indicative conditional "signals robustness with respect to its antecedent" (p. 12) because an indicative conditional that wasn't robust with respect to its antecedent wouldn't be useful for *modus ponens* (p. 13). And if the robustness of a material conditional $p \supset q$ with respect to p is high, then so is the conditional probability of q on p (p. 15). So again, an indicative conditional is assertible only if the probability of its consequent conditional on its antecedent is high.[5]

One problem with the truth-functional approach is that it

> … yields counterintuitive results for sentences containing conditionals as constituents.
>
> For example, it tells us that the following is a tautology:
>
> (If A, B) or (if not-A, B).
>
> So anyone who rejects the first conditional must, on pain of contradiction, accept the second. So if I reject the conditional 'If the Conservatives lose, Thatcher will resign', I am committed to accepting 'If the Conservatives win, Thatcher will resign'!
>
> (Edgington 1991, p. 187)

The essence of the problem here is that we want a theory that handles indicative conditionals well whether or not they are asserted. But theories that are based on conversational implicature do not straightforwardly generalize to non-assertive settings, like the individual disjuncts of a disjunction. While this is a serious problem, recent work in linguistics suggests that it may not be insuperable. That work argues that implicatures can occur in non-assertive settings, and develops explanations of the relevant data (see Chierchia, Fox, and Spector 2012 for a helpful recent survey). No one has yet tried to extend such explanations to conditionals, and there isn't space for an exploration of the possibility here. Moreover, it's very unlikely that such explanations would help with other expressions in the language of subjective uncertainty.

The truth-functional theories sketched in this section aim to establish a link between assertibility and conditional probability. But because these analyses make a sharp distinction between the probability of the proposition expressed by a conditional and the probability of its consequent conditional on its antecedent, they need to explain assertibility in pragmatic terms. In other words, these analyses aim to unite conditional probability and meaning,

[4] Note, however, that Lewis isn't saying (with Adams) that "the probability of an indicative conditional of the form 'if A is the case then B is' is a conditional probability" (1975: p. 3). The next section discusses Lewis's reasons for abandoning the notion of the "probability of an indicative conditional" altogether, and instead talking of "assertibility." For illuminating criticism of Lewis's 1979 approach, see Appiah 1985, pp. 177–81.

[5] Lewis refines Jackson's story somewhat, and endorses the refined story: see the postscript to his 1976.

but aim to effect that unification through an amalgam of pragmatically and semantically conveyed content. This hybrid approach makes it difficult (though perhaps not impossible) to secure compositionality.

35.3 Truth-conditional Semantics and Conditional Probability

Robert Stalnaker's "selection function" analysis is expressly intended to "make the transition from belief conditions to truth conditions; that is, to [provide] a set of truth conditions for statements having conditional form which explains why we use [the Ramsey test] ... to evaluate them" (1968, pp. 44–5; see also his 1970). In particular, Stalnaker attempts to link semantic meaning, in the form of truth conditions, and conditional probability. One aspect of his analysis that is crucial to this aim is the validity of

Conditional excluded middle: Either if a, c, or if a, $\neg c$.[6]

The validity of this principle is intended to reflect the fact that $P(c|a) = 1 - P(\neg c|a)$.

A truth-conditional semantics for conditionals that would explain the connections between conditionals, supposition, and conditional probability would unite a wide range of otherwise distant-looking domains. Alan Hájek writes that if "Stalnaker's Thesis" were true—that is, if it were true that the probability of a conditional equaled the corresponding conditional probability in any probability space that modeled a rational agent's credences[7]—then "logic, probability theory, and Bayesian epistemology would all be enriched" (2011, p. 6). One might add that philosophy of language, formal semantics, and pragmatics would also be enriched, since we can read Stalnaker as trying to integrate formal epistemology—in particular, the dynamics of credal states—into truth-conditional semantics. This is because, as he puts it, the essential innovation in his semantics for conditionals is meant in part to "represent ... methodological policies" governing "how I would revise my beliefs in the face of a particular potential discovery" (Stalnaker 1970, p. 80; cf. Stalnaker 1984, pp. 115–16, 119–20).

But David Lewis's renowned "triviality results," and the many subsequent variations on them, make it notoriously hard to say whether this goal is attainable. Lewis argues that "there is no way to interpret a conditional connective so that, with sufficient generality, the probabilities of conditionals will equal the appropriate conditional probabilities" (1976, p. 135). Of course, little is ever uncontroversial in philosophy: some resist Lewis's results, while others generalize and strengthen them.[8] Here is one perspicuous Lewis-style triviality proof, adapted from Blackburn 1986, p. 219:

[6] Other "modal" analyses, such as C. I. Lewis's strict conditional analysis (1918; Lewis and Langford 1932) and David Lewis's variably strict conditional analysis (Lewis 1973), do not validate this principle, and so do not impute a logic to conditionals that reflects the logic of conditional probability.

[7] As I discuss later, Stalnaker came to reject this thesis; the name for it has nevertheless stuck. It also sometimes goes by "CCCP"—the "conditional construal of conditional probability" (Hájek and Hall 1994) and sometimes by "Adams' Thesis." (Some theorists, however, reserve "Adams' Thesis" for the claim that the *assertibility* of a conditional equals the corresponding conditional probability.)

[8] See section 6 of Edgington 1995 and Hájek 2011 for overviews of the literature.

$$P(a \Rightarrow c) = P(c \wedge (a \Rightarrow c)) + P(\neg c \wedge (a \Rightarrow c)) \tag{35.1}$$
$$= (P((a \Rightarrow c)\,|c) \times P(c)) + (P((a \Rightarrow c)\,|\neg c) \times P(\neg c)) \tag{35.2}$$
$$= (P(c|\,(a \wedge c)) \times P(c)) + (P(c|\,(a \wedge \neg c)) \times P(\neg c)) \tag{35.3}$$
$$= (1 \times P(c)) + (0 \times P(\neg c)) \tag{35.4}$$
$$= P(c) \tag{35.5}$$

For the move from (1) to (2) it suffices that $\frac{P(a \wedge c)}{P(a)} = P(c|a)$, so $P(a \wedge c) = p(c|a) \times P(a)$. For the move from (2) to (3) it suffices that $P((a \Rightarrow c)\,|b) = P(c|a \wedge b)$—intuitively, that there's no difference between the probability of "If b, then if a then c" and "If a and b, then c." But if (5) is true—if the probability of "If a, then c" just is the probability of c—then the probability of the conditional is completely trivialized. Results of this ilk led Stalnaker to reject "Stalnaker's Thesis,"[9] and led Lewis to endorse the theory of indicative conditionals sketched in the last section.

Here is another important problem for the project of integrating probability theory into truth-conditional semantics, due to Allan Gibbard. Suppose that

> Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says "If Pete called, he won," and Jack slips me a note which says "If Pete called, he lost." I know that these notes both come from my trusted henchmen, but do not know which of them sent which note. I conclude that Pete folded.
>
> (1981, p. 231)[10]

As Gibbard points out, analyses of conditionals like Stalnaker's "share a law of *Conditional Non-contradiction*: that $a \rightarrow \overline{b}$" is inconsistent with $a \rightarrow b$" (p. 231). But if Zack and Jack have contradicted each other, then one of them must have said something false. Intuitively, though, neither has said anything false; indeed the formation of the beliefs they express on their notes looks impeccable. Gibbard concludes that to save analyses like Stalnaker's we must posit "radical dependence" of the semantic value of conditionals on the "utterer's epistemic state" (p. 234)—something that "the audience does not know" (p. 233). Stalnaker (1984, pp. 109–11) and others—including the influential linguist Angelika Kratzer (1986, p. 655) and many influenced by her—embrace this kind of context-sensitivity and try to explain how we can accept that

> conditionals [are] too closely tied to the epistemic states of the agents who utter them for those conditionals to express propositions which could be separated from the contexts in which they are accepted
>
> (Stalnaker 1984, p. 111).[11]

[9] See Rothschild 2013 for an interesting recent attempt to save a restricted form of Stalnaker's Thesis.

[10] Edgington 1997: p. 107 gives a nice way to set up Gibbard cases in which the people playing the roles of Zack and Jack are epistemically on a par.

[11] See especially Stalnaker 2011: pp. 242–4; see also his 2005.

A full assessment of analyses like these is beyond the scope of this chapter. Suffice it to say that for conditionals, at least, the goal of tightly connecting probability and truth-conditional semantics is not obviously attainable. We should see what would happen if we were to leave truth conditions behind, thinking of the vehicles of communication in some other way.

35.4 Non-compositional Non-truth-conditional Theories

On Dorothy Edgington's "conditional assertion" analysis, the indicative conditional "'If *A*, *B*' is an assertion of *B* when *A* is true, and an assertion of nothing when *A* is false" (1995: p. 290).[12] Whether a conditional has a truth value, according to Edgington, depends at least in part on whether its antecedent is true: "my conditional assertion is true if *A* and *B* are both true, and false if *A* is true and *B* is not, and has no truth value when *A* is false." She holds that belief in a conditional is a "conditional belief" and hence "not belief that something is true" and also "not belief that [something] is not false." Rather, "Belief that if *A*, *B* is a conditional belief that it is true given that it has a truth value" (p. 290). But the degree of belief in the closely related belief "that 'If *A*, *B*' is true, given that it has a truth value, is just $b(A\&B)/b(A)$," vindicating a version of the thesis that one's degree of belief in *B* if *A* equals one's conditional probability of *B* on *A* (pp. 290, 262–3).

Another influential non-truth-conditional theory of indicative conditionals goes back to Ernest Adams' (1965), (1966), and (1975) (see also Jeffrey 1964 and Ellis 1973). While Adams makes some allusions to conditional assertion (see especially 1965: p. 196), conditional probability is foremost in his theory, and Adams doesn't flesh out any connections between the two notions. Adams sidesteps Lewis's triviality results by holding that conditionals are sometimes neither true nor false (Adams 1975: p. 7): "it is hopeless to hunt for the 'right' truth conditions for conditionals … if it is also required that truth-conditional soundness should closely approximate probabilistic soundness" (1975: p. 5). But Adams doesn't aspire to give a general compositional semantics for indicative conditionals. He focuses instead on patterns of reasoning, motivation, and action. Indeed, he sees his approach as "largely independent" of concerns about "Speech Acts, … language and communication" (1975: p. 71).

Aspects of Adams' theory were taken up by many theorists more directly concerned with natural language, however. Gibbard, for example, writes that any account of communication based on mutual intention recognition, in the spirit of Grice (1957), "will extend naturally to communication of conditional belief" without requiring that "conditional beliefs must be communicated by means of conditional propositions."

> In felicitous cases, I utter an indicative conditional, and thereby insure that the audience comes to accept that I have a certain conditional belief, belief in *b* given *a*. The audience does so because it trusts my sincerity and command of language. The audience then infers

[12] For a fascinating history of conditional assertion views, see Milne 1997: pp. 197–212. See also Edgington 2008, which develops a conditional assertion account of counterfactuals.

> from my believing *b* given *a* that I have some good grounds for so believing, and takes that as a reason for itself believing *b* given *a*.
>
> (1981, p. 230)

Note that on Gibbard's account, the audience takes the speaker's utterance as evidence that the speaker has good grounds for believing *b* given *a*. To the extent that the audience believes the speaker is epistemically well-placed, the audience then has grounds for believing *b* given *a*. Such views are sometimes called "expressivist," on the grounds that the force of the speaker's utterance is an *expression* of her own conditional belief. If the utterance secures "uptake" in the audience, that is in part thanks to the audience making inferences about the speaker's grounds for believing what she expresses. Note also that the "command of language" essential to Gibbard's explanation is simply that the speaker and audience associate the conditional belief in *b* given *a* with "If *a*, *b*."

What happens when a conditional occurs in an embedded linguistic context?

> Either we need new semantic rules for many familiar connectives and operators when applied to indicative conditionals—perhaps rules of truth, perhaps special rules of assertability like the rule for conditionals themselves—or else we need to explain away all seeming examples of compound sentences with conditional constituents.
>
> (Lewis 1976, p. 142)

Gibbard takes the second horn of this dilemma, arguing that many embeddings of conditionals don't make sense, and that every embedding that *does* make sense is "explainable in an *ad hoc* way" (1981, p. 238). For example, Gibbard argues that sometimes indicative conditionals have an "*obvious basis*: a proposition *c* such that it is presupposed, for both utterer and audience, that [the utterer] will believe the consequent given the antecedent iff he believes *c*" (p. 237). In such cases, Gibbard suggests, we interpret the embedded conditional as if it simply expressed its "obvious basis." For example, we might interpret

(7) If the cup broke if dropped, then it was fragile.

as

(8) If the cup was disposed to break on being dropped, then it was fragile.

Gibbard and other advocates of non-truth-conditional analyses are clear that they do not have a general theory of embedding to offer (Adams 1965; Edgington 1995, p. 284). But Gibbard suggests that if ad hoc explanations suffice where embedding *is* possible, then this is not a significant cost.

Such ad hoc explanations—and the need to appeal to them in the first place—leave most linguists and an increasing number of philosophers of language dissatisfied. For example, the influential semanticist Kai von Fintel writes that accounts like Adams', Gibbard's, and Edgington's have had "no impact at all in linguistic work on natural language semantics" (2011, p. 1524). The lack of impact is due to the fact that such accounts abandon compositionality, and so need *lots* of ad hoc explanations. Such explanations are often very difficult to assess. For example, Edgington discusses (9), which has a quantifier scoped over a conditional, at length:

(9) There is a boy in my class who, if I criticize him, will get angry.

(Kölbel 2000, p. 105)

After considering and rejecting several ways in which one might paraphrase (9), she writes that "we are free to construe it . . . as saying something along the following lines":

(10) There is a boy in the class such that, on the supposition that I criticize him, he will get angry.

(2000, p. 114)

Suppose for the sake of argument that this is a successful paraphrase of (9). What has been gained by offering it? Semanticists would say that the paraphrase does nothing to help us analyze (9) unless we have a detailed analysis of (10) itself. And while Edgington does describe credal states that she thinks typically accompany sincere assertions of (9), this is not to offer a semantic value for (9), let alone for the expressions in it. For her part, Edgington would respond that the meaning of "if" is not explained via semantic value, but rather by the respects in which conditional assertion differs from assertion simpliciter. It would be helpful, though, to know more about how speech acts and semantic values are supposed to interact in (9). Do speakers use (9) to make a conditional assertion? One might think not; (9) and its putative paraphrase seem to be used to assert something about a boy in the class. But then it seems that there isn't anything distinctive about the conditional assertion account of sentences like (9), leaving us no better off than we were with truth-conditional analyses.

Stepping back a bit: according to conditional assertion accounts, what sort of vehicle of communication is associated with a conditional varies depending on whether the antecedent of the conditional is true. That's what it is for an assertion to be conditional: the assertion of the consequent is made if and only if the antecedent is true; otherwise we have "an assertion of nothing" (Edgington 1995, p. 290). The act of uttering a conditional can be helpful to communication whether or not the antecedent is true, however.

> I say to you "If you press that switch, there will be an explosion". As a consequence, you don't press it. Had I said nothing at all, let us suppose you would have pressed it. A disaster is avoided, as a result of this piece of linguistic communication. It is not as if nothing had been said. This is no objection to the idea that I did not (categorically) assert anything. For let us suppose that I am understood as having made a conditional assertion of the consequent. My hearer understands that if she presses it, my assertion of the consequent has categorical force; and, given that she takes me to be trustworthy and reliable, if it does acquire categorical force, it is much more likely to be true than false. So she too acquires reason to think that there will be an explosion if she presses it, and hence a reason not to press it.
>
> (Edgington 1995, p. 289)

Like truth-functional theories, then, Edgington's theory treats the vehicles of communication as an amalgam of semantically conveyed and pragmatically conveyed content. As with truth-functional analyses, compositionality is a challenge for this sort of approach but not clearly an insuperable challenge. I think it is fair to say, however, that the prospects for making conditional assertion analyses compositional are dimmer than they are for truth-functional analyses. This is because the case for grammaticalized speech acts is much weaker than the case for grammaticalized implicatures.

35.5 Compositional Non-truth-conditional Theories

The last option to consider is the development of compositional semantic theories delivering vehicles of communication that, unlike truth conditions, directly represent probabilities and relationships between probabilities. But while conditionals help motivate this approach—and are, again, perhaps the most historically important motivation for it—they are just one motivation of many. After all, we also want to analyze sentences like our earlier

(1) There's a 4% chance that Bob won the raffle.

So in developing a positive account, it's important to step back and think about the language of subjective uncertainty in a more general way. This perspective will help us explore ways of theorizing about conditionals that generalize to other expressions of subjective uncertainty.

It's also important to see that a dialectic similar to the one discussed for conditionals applies to the explicitly quantitative parts of the language of subjective uncertainty, even though its development in the literature is less extensive and more recent. Suppose, for example, that we tried to give a truth-conditional analysis of a sentence such as (1). We might treat it as elliptical for (11):

(11) I believe that there's a 4% chance that Bob won the raffle.

But this approach seems to attribute the wrong subject matter to (1)—(1) simply isn't about the speaker's psychological state (Bennett 2003, p. 90; see also Yalcin 2011). It also seems to mischaracterize the intended effect of (1), leading to the prediction that I can believe you when you assert (1), while being sure that Bob didn't win the raffle. It's more promising, I think, to appeal to a less subjective notion of evidential probability—perhaps that developed in Williamson (2000)—and then to analyze (1) as elliptical for

(12) The evidential probability that Bob won the raffle is 4%.

Such accounts are plausible only if typical speakers have the appropriate kind of epistemic access to evidential probabilities in any situation in which they can appropriately use the language of subjective uncertainty. I'm skeptical of the thought that they do, but there is insufficient space here for an in-depth discussion of the issue.

Another kind of account takes the operator "There's a 4% chance that" in (1) to indicate that the way in which the speaker puts forward the proposition that Bob won the raffle is attenuated.[13] Such "force modifier" approaches have had many advocates in philosophy and in linguistics over the years, but like the non-truth-conditional theories considered earlier, they have trouble with embedding. It is particularly difficult to see how to generalize force modifier accounts to Quine's "third grade of modal involvement" (1953): sentences in which a quantifier takes scope over a modal operator. In this case, it is difficult to see how

[13] See, e.g., Toulmin 1950, Lyons 1977, Forrest 1981, Price 1983a and 1983b, and Yalcin 2005; see Swanson 2010 for more citations.

force modifier accounts can handle a quantifier scoped over an expression of subjective uncertainty (for discussion, see Swanson 2006, p. 61; and Swanson 2010).

In sum, truth-conditional analyses of the quantitative language of subjective uncertainty look problematic, and so do non-compositional analyses. The remaining corner of logical space—compositional analyses that aren't truth-conditional—includes many different kinds of analyses. Some, like dynamic semantics, have much to say about conditionals, but at this point there is little work tying such approaches to probability.[14] For this reason, and because space is limited, I move directly to a compositional approach that is designed from the outset to interface well with probability.

The interpretation function of constraint semantics takes declarative sentences not to propositions but to *constraints* (Swanson 2006 and 2012; Yalcin 2007 and 2010; Moss 2015). Intuitively, a constraint is a characterization of states an addressee could be in that are compatible with a sentence. For example, if we did not want to be realists about tastiness, we might say that the constraint associated with "Artichoke hearts are tastier than broccoli" is the set of gustatory preferences according to which artichoke hearts are ranked higher than broccoli. On this application of the constraint semantic framework, when a speaker says that artichoke hearts are tastier than broccoli, she advises her addressees to have gustatory preferences that conform to the constraint that artichoke hearts are ranked higher than broccoli. To effect the move to constraints compositionally, other semantic types are changed as well. For example, in an intensional semantic theory, the meaning of a predicate like "is tall" is often modeled as a function that takes an individual concept and yields a proposition. In an intensional constraint semantics, the meaning of "is tall" would be modeled as a function that takes an individual concept and yields a constraint (Swanson 2006, pp. 50–51).

When we apply constraint semantics to the language of subjective uncertainty, we need constraints on credal states. For example, the constraint associated with

(1) There's a 4% chance that Bob won the raffle.

is the set of credal states that assign credence 0.04 to the proposition that Bob won the raffle. To say that Alice believes that there is a 4% chance that Bob won the raffle is to say that Alice's credal state is an element of the constraint associated with (12). This kind of approach sets itself apart from "force modifier" accounts by securing compositionality below the clausal level. But to do this, constraint semantics must appeal to a function that takes a set of constraints and yields "the constraint associated with the disjunction of sentences that express those constraints" (Swanson 2012, p. 14). The precise characterization of this function is not a project for semantics or philosophy of language—just as, say, the precise characterization of the semantic value of "justice" is not a project for semantics or philosophy of language. Rather, just as ethicists and social and political philosophers work to improve our understanding of justice, formal epistemologists work toward improving our understanding of the functions that constraint semantics should deploy. What's important to see, on the side of philosophy of language, is that a compositional semantic theory can *accommodate* a wide range of characterizations of this function and of other functions that are important to the language of subjective uncertainty.

[14] Yalcin 2012 starts to develop such connections.

With this overall picture in mind, let's return to our discussion of conditionals. While I do not mean to take a stand on whether this is the *right* analysis of simple conditionals such as (4), constraint semantics can easily accommodate the thought that the meaning of such conditionals is tightly tied to conditional probability.

(4) If Oswald did not kill Kennedy, then someone else did.

It is straightforward, for example, to write a constraint semantic entry for "if" on which the semantic value of (4) is the set of probability measures in which the probability of someone else's having killed Kennedy, conditional on Oswald's not having killed Kennedy, is 1 (Swanson 2006, pp. 67–9).[15] A theorist who thinks conditional probability comes apart from belief in conditionals—a theorist convinced, for example, by the relevant arguments in McGee 2000 or Kaufmann 2004—would give "if" a different constraint semantic entry. Arguably there is even a sense in which we would want *both* semantic entries: one for highly idealized language users, who are in particular unusually adept conditionalizers, and another designed to match the judgments of most actual language users. But once the favored characterization of the circumstances in which one believes a conditional is complete, we can write a constraint semantic entry to match that characterization. Similarly, it isn't obvious what constraint should be associated with embeddings of conditionals. Many have despaired, for example, in the face of Gibbard's

(13) If Kripke was there if Strawson was, then Anscombe was there.

Gibbard asks us to imagine being told this "of a conference you don't know much about," and asks, rhetorically, "Do you know what you have been told?" (1981, p. 235).[16] But because constraint semantics can accommodate many different characterizations of the semantic contribution of "if", we are free to explore a wide range of possible characterizations and see which work best.

In a slogan: the need for a compositional semantic theory is neither a bar nor a guide to the task of characterizing complex credal and doxastic states. This slogan does exaggerate a bit. A simpler or more elegant semantic entry might be preferable to another entry for some purposes—for example, for modeling *how* humans understand language, as opposed to representing linguistic competence at a more abstract level. Moreover, finding the right roles for semantics and pragmatics (and the right balance between them) will eventually be

[15] One such entry would "reverse engineer" the propositions targeted by the constraints associated with the conditional's antecedent and consequent, respectively. Then it would yield the constraint according to which the probability of the proposition associated with the antecedent is equal to the probability of the conjunction of that proposition and the proposition associated with the consequent, thus making the probability of the consequent conditional on the antecedent 1. It's also possible to give a less demanding analysis, on which the relevant conditional probability must meet or exceed some threshold lower than 1.

[16] This conditional is doubtless hard to interpret in many contexts. To see that it is interpretable, imagine that Kripke goes to a conference that Strawson goes to only if Kripke thinks it's an important conference. Then (13) might be used to say that Anscombe goes to every conference that Kripke thinks is important. For an interesting recent discussion of embedded conditionals, see Sennet and Weisberg 2012.

important as well. But much work in formal epistemology is necessary first. The payoff of incorporating such work into the constraint semantic framework would be a compositional semantics for a natural language that would connect the semantic value of declarative sentences to credal states in a very direct way.

35.6 Conclusion

The approaches surveyed here differ greatly in how they see the relationship between philosophy of language and linguistics, on the one hand, and formal epistemology on the other. Non-truth-conditional theories aim to make the vehicles of communication resemble the models of credal states developed by formal epistemologists. Truth-conditional theories don't have this aim. And cross-cutting the distinction between truth-conditional and non-truth-conditional approaches, different theorists put differing emphases on the importance of compositionality.

The door is not shut on any of these approaches. But there has been a recent push toward compositional non-truth-conditional theories of the language of subjective uncertainty. Many factors have contributed to this push, including the recent surge of interest in epistemic modals (see the papers in Egan and Weatherson 2011 for an overview). But two factors are especially likely to make this trend last. First, philosophers of language are increasingly aware of and concerned by the phenomena that make compositionality so important to semanticists and other linguists. Secondly, philosophers of language increasingly aspire to explain data that seem to demand collaborative work with formal epistemologists. This kind of work is facilitated by frameworks that, like constraint semantics, allow the representations of credal and doxastic states developed by formal epistemologists to be incorporated into compositional semantics.

Successes in this framework would enrich our perspective on the questions about the general features of language with which we began. If we think of linguistic communication as fundamentally a matter of describing the world, then it is natural to think that the vehicles of communication must be truth-apt. But the difficulties involved with analyzing the language of subjective uncertainty give us some reason instead to think of linguistic communication as a way of advising others about features of our perspective on the world.

Acknowledgments

For helpful comments, thanks to John Cusbert and Alan Hájek.

References

Adams, E. W. (1965) The Logic of Conditionals. *Inquiry*. 8. pp. 166–97.

Adams, E. W. (1966) Probability and the Logic of Conditionals. *Studies in Logic and the Foundations of Mathematics*. 43. pp. 256–316.

Adams, E. W. (1970) Subjunctive and Indicative Conditionals. *Foundations of Language*. 6. pp. 89–94.
Adams, E. W. (1975) *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: Reidel.
Appiah, A. (1985) *Assertion and Conditionals*. Cambridge: Cambridge University Press.
Austin, J. L. (1962) *How to Do Things with Words*. London: Oxford University Press.
Bennett, J. (2003) *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.
Blackburn, S. (1986) How Can We Tell Whether a Commitment Has a Truth Condition? In Travis, C. (ed.) *Meaning and Interpretation*. pp. 201–32. Oxford: Blackwell.
Chierchia, G., Fox, D., and Spector, B. (2012) Scalar Implicature as a Grammatical Phenomenon. In Maienborn, C., von Heusinger, K., and Portner, P. (eds.) *Semantics: An International Handbook of Natural Language Meaning*. Vol. 3. pp. 297–331. Berlin: Mouton de Gruyter.
Edgington, D. (1991) Do Conditionals have Truth-Conditions? In Jackson, F. (ed.) Conditionals. pp. 176–201. Oxford: Oxford University Press.
Edgington, D. (1995) On Conditionals. *Mind*. 104. pp. 235–329.
Edgington, D. (1997) Truth, Objectivity, Counterfactuals and Gibbard. *Mind*. 106. pp. 107–16.
Edgington, D. (2000) General Conditional Statements: A Response to Kölbel. *Mind*. 109. pp. 109–16.
Edgington, D. (2008) Counterfactuals. *Proceedings of the Aristotelian Society*. 108. pp. 1–21.
Egan, A. and Weatherson, B. (2011) *Epistemic Modality*. Oxford: Oxford University Press.
Ellis, B. (1973) The Logic of Subjective Probability. *British Journal for the Philosophy of Science*. 24. 2. pp. 125–52.
Forrest, P. (1981) Probabilistic Modal Inferences. *Australasian Journal of Philosophy*. 59. 1. pp. 38–53.
Gibbard, A. (1981) Two Recent Theories of Conditionals. In Harper, W., Stalnaker, R., and Pearce, G. (eds.) *Ifs*. pp. 211–47. Dordrecht: Reidel.
Gillies, A. (2009) On Truth-Conditions for *If* (but Not Quite Only *If*). *Philosophical Review*. 118. 3. pp. 325–49.
Grice, P. (1957) Meaning. In *Studies in the Way of Words*. pp. 213–23. Cambridge, MA: Harvard University Press.
Grice, P. (1987) Logic and Conversation. pp. 22–40. In *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Hájek, A. (2011) Triviality Pursuit. *Topoi*. 30. pp. 3–15.
Hájek, A. and Hall, N. (1994) The Hypothesis of the Conditional Construal of Conditional Probability. In Eells, E. and Skyrms, B. (eds.) *Probability and Conditionals: Belief Revision and Rational Decision*. pp. 75–111. Cambridge: Cambridge University Press.
Jackson, Frank (1979) On Assertion and Indicative Conditionals. In *Mind, Method and Conditionals: Selected Essays*. pp. 3–26. New York, NY: Routledge.
Jeffrey, Richard C. (1964) If. *Journal of Philosophy*. 61. pp. 702–3.
Jeffrey, Richard C. (1986) Probabilism and Induction. *Topoi*. 5. pp. 51–8.
Joyce, James M. (1998) A Nonpragmatic Vindication of Probabilism. *Philosophy of Science*. 65. 4. pp. 575–603.
Joyce, James M. (2005) How Probabilities Reflect Evidence. *Philosophical Perspectives*. 19. 1. pp. 153–78. doi:10.1111/j.1520-8583.2005.00058.x.
Kaufmann, Stefan (2004) Conditioning Against the Grain: Abduction and Indicative Conditionals. *Journal of Philosophical Logic*. 33. pp. 583–606.

Kölbel, Max (2000) Edgington on Compounds of Conditionals. *Mind*. 109. 433. pp. 97–108.
Kratzer, Angelika (1986) Conditionals. In von Stechow, A. and Wunderlich, D. (eds.) Semantics: An International Handbook of Contemporary Research. pp. 651–656. Berlin: W. de Gruyter.
Lewis, C. I. (1918) *A Survey of Symbolic Logic*. Berkeley, CA: University of California Press.
Lewis, C. I. and Langford, C. H. (1932) *Symbolic Logic*. New York, NY: Century Company.
Lewis, David K. (1973) *Counterfactuals*. Malden, MA: Basil Blackwell Ltd.
Lewis, David K. (1976) Probabilities of Conditionals and Conditional Probabilities. In *Philosophical Papers*. Vol. 2. pp. 133–56. Oxford: Oxford University Press.
Lyons, John (1977) *Semantics*. Vol. 2. Cambridge: Cambridge University Press.
McGee, Vann (2000) To Tell the Truth about Conditionals. *Analysis*. 60. 1. pp. 107–11.
Milne, Peter (1997) Bruno de Finetti and the Logic of Conditional Events. *British Journal for the Philosophy of Science*. 48. pp. 195–232.
Moss, Sarah (2015) On the Semantics and Pragmatics of Epistemic Vocabulary. *Semantics and Pragmatics*. 8. 5. pp. 1–81.
Price, Huw (1983a) 'Could a Question Be True?': Assent and the Basis of Meaning. *Philosophical Quarterly*. 33. 133. pp. 354–64.
Price, Huw (1983b) Does 'Probably' Modify Sense? *Australasian Journal of Philosophy*. 61. 4. pp. 396–408.
Quine, W. V. O. (1953) Three Grades of Modal Involvement. In *The Ways of Paradox*, pp. 158–76. New York, NY: Random House.
Ramsey, Frank Plumpton (1931) General Propositions and Causality. In Mellor, D. H. (ed.) *Philosophical Papers*. pp. 145–63. Cambridge: Cambridge University Press.
Rothschild, Daniel (2013) Do Indicative Conditionals Express Propositions? *Noûs*. 47. 1. pp. 49–68.
Sennet, Adam and Weisberg, Jonathan (2012) Embedding 'If and Only If.' *Journal of Philosophical Logic*. 41. 2. pp. 449–60.
Stalnaker, Robert C. (1968) A Theory of Conditionals. In Harper, W., Stalnaker, R., and Pearce, G. (eds.) *Ifs*. pp. 41–55. Dordrecht: Reidel.
Stalnaker, Robert C. (1970) Probability and Conditionals. *Philosophy of Science*. 37. 1. pp. 64–80.
Stalnaker, Robert C. (1984) *Inquiry*. Cambridge, MA: MIT Press.
Stalnaker, Robert C. (1986) Possible Worlds and Situations. *Journal of Philosophical Logic*. 15. pp. 109–23.
Stalnaker, Robert C. (2005) Conditional Assertions and Conditional Propositions. In Gajewski, J., Hacquard, V., Nickel, B., and Yalcin, S. *New Work on Modality*. MIT Working Papers in Linguistics. Vol. 51. Cambridge, MA: MIT Press.
Stalnaker, Robert C. (2011) Conditional Propositions and Conditional Assertions. In Egan, A. and Weatherson, B. (eds.) *Epistemic Modality*. pp. 227–48. Oxford: Oxford University Press.
Swanson, Eric (2006) *Interactions with Context*. Cambridge, MA: MIT Press.
Swanson, Eric. (2010) On Scope Relations between Quantifiers and Epistemic Modals. *Journal of Semantics*. 27. 4. pp. 529–40.
Swanson, Eric (2016) The Application of Constraint Semantics to the Language of Subjective Uncertainty. *Journal of Philosophical Logic*. 45. 2. pp. 121–146.
Thomson, James F. (1990) In Defense of '⊃'. *Journal of Philosophy*. 87. 2. pp. 57–70.

Toulmin, S. E. (1950) Probability. *Proceedings of the Aristotelian Society*. Supplementary Volumes. 24. pp. 27–62.

Von Fintel, Kai (2011) Conditionals. In von Heusinger, Klaus, Maienborn, Claudia, and Portner, Paul (eds.) *Semantics: An International Handbook of Meaning*. Vol. 2. pp. 1515–38. Berlin/Boston: Mouton de Gruyter.

Williamson, Timothy (2000) *Knowledge and Its Limits*. Oxford: Oxford University Press.

Yalcin, Seth (2005) Epistemic Modals. In Gajewski, J. Hacquard, V., Nickel, B, and Yalcin, S. (eds.) *New Work on Modality*. pp. 231–72. MIT Working Papers in Linguistics. Vol. 51. Cambridge, MA: MIT Press.

Yalcin, Seth (2007) Epistemic Modals. *Mind*. 116. pp. 983–1026.

Yalcin, Seth (2010) Probability Operators. *Philosophy Compass*. 5. 11. pp. 916–37.

Yalcin, Seth (2011) Nonfactualism about Epistemic Modality. In Egan, A. and Weatherson, B. (eds.) *Epistemic Modality*. Oxford: Oxford University Press.

Yalcin, Seth (2012) Context Probabilism. In *Proceedings of the Eighteenth Amsterdam Colloquium*. Lecture Notes in Computer Science. Volume 7218. pp. 12–21. Springer.

Yalcin, Seth (2014) Semantics and Metasemantics in the Context of Generative Grammar. In Burgess, A. and Sherman, B. (eds.) *Metasemantics: New Essays on the Foundations of Meaning*. pp. 17–54. Oxford: Oxford University Press.

CHAPTER 36

DECISION THEORY

LARA BUCHAK

36.1 Introduction

Decision theory has at its core a set of mathematical theorems that connect rational preferences to functions with certain structural properties. The components of these theorems, as well as their bearing on questions surrounding rationality, can be interpreted in a variety of ways. Philosophy's current interest in decision theory represents a convergence of two very different lines of thought, one concerned with the question of how one ought to act, and the other concerned with the question of what action consists in and what it reveals about the actor's mental states. As a result, the theory has come to have two different uses in philosophy, which we might call the *normative* use and the *interpretive* use. It also has a related use that is largely within the domain of psychology, the *descriptive* use.

The first two sections of chapter examine the historical development of normative decision theory and the range of current interpretations of the elements of the theory, while the third section explores how modern normative decision theory is supposed to capture the notion of rationality. The fourth section presents a history of interpretive decision theory, and the fifth section examines a problem that both uses of decision theory face. The sixth section explains the third use of decision theory, the descriptive use. Section seven considers the relationship between the three uses of decision theory. Finally, section eight examines some modifications to the standard theory and the conclusion includes some remarks about how we ought to think about the decision-theoretic project in light of a proliferation of theories.

36.2 Normative Decision Theory

The first formal decision theory was developed by Blaise Pascal in correspondence with Pierre Fermat about "the problem of the points," the problem of how to divide up the stakes

of players involved in a game if the game ends prematurely.[1] Pascal proposed that each gambler should be given as his share of the pot the monetary expectation of his stake, and this proposal can be generalized to other contexts: the monetary value of a risky prospect is equal to the expected value of that prospect. Formally, if $L = \{\$x_1, p_1; \$x_2, p_2; \ldots\}$ represents a "lottery" which yields $\$x_i$ with probability p_i, then its value is:

$$EV(L) = \sum_{i=1}^{\infty} p_i x_i$$

This equivalence underlies a prescription: when faced with two lotteries, you ought to prefer the lottery with the higher expected value, and to be indifferent if they have the same expected value. More generally, you ought to *maximize expected value*.

This norm is attractive for a number of reasons. For one, it enjoins you to make the choice that would be better over the long run if repeated: over the long run, repeated trials of a gamble will average out to their expected value. For another, going back to the problem of the points, it ensures that players will be indifferent between continuing the game and leaving with their share. But there are several things to be said against the prescription. One is that it is easy to generate a lottery whose expected value is infinite, as shown by the St. Petersburg Paradox (first proposed by Nicolas Bernoulli). Under the norm in question, one ought to be willing to pay any finite amount of money for the lottery {$1, 1/2; $2, 1/4; $4, 1/8; …}, but most people think that the value of this lottery should be considerably less. A second problem is that the prescription does not seem to account for the fact that whether one should take a gamble depends on what one's total fortune is: one ought not risk one's last dollar for an even chance at $2, if losing the dollar means that one will be unable to eat. Finally, the prescription doesn't seem to adequately account for the phenomenon of risk aversion: most people would rather have a sure thing sum of $x than a gamble whose expectation is $x (for example, $100 rather than {$50, 1/2; $150, 1/2}) and don't thereby seem irrational.

In response to these problems, Daniel Bernoulli (1738/1954) and Gabriel Cramer (see Bernouilli 1738/1954: 33) each independently noted that the satisfaction that money brings diminishes the more money one has, and each proposed that the quantity whose expectation one ought to maximize is not money itself but rather the "utility" of one's total wealth. (Note that for Bernoulli, the outcomes are total amounts of wealth rather than changes in wealth, as they were for Pascal.) Bernoulli proposed that an individual's utility function of total wealth is $u(\$x) = \log(x)$. Therefore, the new prescription is to maximize:

$$EU(L) = \sum_{i=1}^{\infty} p_i u(\$x_i) = \sum_{i=1}^{\infty} p_i log(x_i)$$

This guarantees that the St. Petersburg lottery is worth a finite amount of money; that a gamble is worth a larger amount of one's money the wealthier one is; and that the expected utility of any lottery is less than the utility of its monetary expectation.

Notice that the norm associated with this proposal is objective in two ways: it takes the probabilities as given, and it assumes that everyone should maximize the same utility

[1] See Fermat and Pascal (1654/1929).

function. One might reasonably wonder, however, whether everyone does get the same amount of satisfaction from various amounts of money. Furthermore, the proposal tells us nothing about how we ought to value lotteries with non-monetary outcomes, which are also plausibly of different value to different people. A natural thought is to revise the norm to require that one maximize the expectation of one's own, *subjective* utility function, and to allow that the utility function take any outcome as input.

The problem with this thought is that it is not clear that individuals have access to their precise utility functions through introspection. Happily, it turns out that we can implement the proposal without such introspection: John von Neumann and Oskar Morgenstern (1944) discovered a *representation theorem* that allows us to determine merely from an agent's pair-wise preferences whether she is maximizing expected utility and if she is, allows us to determine an agent's entire utility function from these preferences. Von Neumann and Morgenstern identified a set of axioms on preferences over lotteries such that if an individual's preferences conform to these axioms, then there exists a utility function of outcomes, unique up to positive affine transformation, that represents her as an expected utility maximizer.[2] The utility function represents her in the following sense: for all lotteries L_1 and L_2, the agent weakly prefers L_1 to L_2 if and only if L_1 has at least as high an expected utility as L_2 according to the function. Thus, we can replace expected objective utility maximization with expected subjective utility maximization as an implementable norm, even if an agent's utility function is opaque to her.

Leonard Savage's (1954/1974) representation theorem took the theory one step further. Like von Neumann and Morgenstern, Savage allowed that an individual's values were up to her. But Savage was primarily interested not in how an agent should choose between lotteries when she is given the exact probabilities of outcomes, but rather in how an agent should choose between ordinary acts when she is uncertain about some feature of the world: for example, how she should choose between breaking a sixth egg into her omelet and refraining from doing so, when she does not know whether or not the egg is rotten. Savage noted that an act leads to different outcomes under different circumstances, and, taking an outcome to be specified so as to include everything an agent cares about, he defined the technical notion of an act as a function from possible states of the world to outcomes.[3] For example, the act of breaking the egg is the function {egg is good → I eat a 6-egg omelet; egg is rotten → I throw away the omelet}. More generally, we can represent an act f as $\{E_1, x_1; \ldots; E_n, x_n\}$, where E_i are mutually exclusive and exhaustive events (an event being a set of states), and each state in E_i results in outcome x_i under act f.[4] Savage's representation theorem shows that an agent's preferences over these acts suffice to determine both her subjective utility function of outcomes and her subjective probability function of events, provided her pair-wise preferences conform to the axioms of his theorem.[5] Formally, u and p represent an

[2] I will often talk about an agent's utility function when strictly speaking I mean the family of utility functions that represents her. However, facts about the utility function that are not preserved under affine transformation, such as the zero point, will not count as "real" facts about the agent's utility values.

[3] Savage used the terminology "consequence" where I am using "outcome."

[4] Savage also treats the case in which the number of possible outcomes of an act is not finite (Savage (1954/1974: pp. 76–82), although his treatment requires bounding the utility function. Assuming each act has a finite number of outcomes will simplify the discussion.

[5] Again, the utility function is unique up to positive affine transformation. The probability function is unique.

agent's preferences if and only if she prefers the act with the highest expected utility, relative to these two functions:

$$EU\left(f\right) = \sum_{i=1}^{n} p\left(E_i\right) u\left(x_i\right)$$

Savage's theory therefore allows that both the probability function and the utility function are subjective. The accompanying prescription is to maximize expected utility, relative to these two functions.

Since Savage, other representation theorems for subjective expected utility theory have been proposed, most of which are meant to respond to some supposed philosophical problem with Savage's theory.[6] One set of issues surrounds what we should prefer when utility is unbounded and acts can have an infinite number of different outcomes, or when outcomes can have infinite utility value.[7] Another set of issues concerns exactly what entities are the relevant ones to assign utility and probability to in decision-making. The developments in this area begin with Richard Jeffrey (1965), who objected to Savage's separation between states, outcomes, and acts, and argued that the same objects ought to be the carriers of both probability and value. Jeffrey proposed a theory on which both the probability and utility function take propositions as inputs. Axiomatized by Ethan Bolker (see Jeffrey 1965: pp. 142–3, 149), Jeffrey's theory enjoins the agent to maximize:[8]

$$u\left(A\right) = \sum_{i=1}^{\infty} p\left(S_i|A\right) u\left(S_i \,\&\, A\right)$$

where S_i and A both stand for arbitrary propositions (they range over the same set), but S_i is to play the role of a state and A of an act. Bolker's (1965-67) representation theorem provides axioms on a preference relation over the set of propositions that allow us to extract p and u, although the uniqueness conditions are more relaxed than in the aforementioned theories. Jeffrey proposed that we ought to interpret the items that an agent has preferences over as "news items"; so, for example, one is asked whether one would prefer the news that one breaks the egg into one's omelet or that one does not. The connection to action, of course, is that one has the ability to create the news when it comes to propositions about acts one is deciding between.

Certain features of Jeffrey's interpretation are inessential to the maximization equation. It is not necessary to follow Jeffrey in interpreting preferences as being about news items. Nor is there consensus that p and u ought to have as their domain the same set of objects.[9] For example, while it is clear that we can assign utility values to acts under our own control, Wolfgang Spohn (1977) and Isaac Levi (1991) each argue that we cannot assign probability values to these acts.

[6] See Fishburn (1981) for a helpful catalogue of some of these.

[7] See, for example, Vallentyne (1993), Nover and Hájek (2004), Bartha (2007), Colyvan (2008), and Easwaran (2008).

[8] Jeffrey used a slightly different, but equivalent, formulation. He also used functions named *prob* and *des* rather than p and u, but the difference is terminological.

[9] While this feature is inessential to Jeffrey's maximization equation as written above, it is essential to Bolker's representation theorem.

Another issue with Jeffrey's theory has been the source of a significant development in decision theory. Because the belief component of Jeffrey's theory corresponds to conditional probabilities of states given acts, this component will have the same numerical value whether an act causes a particular outcome or is merely correlated with it. Therefore, agents will rank acts that are merely correlated with preferred outcomes the same as acts that tend to cause preferred outcomes. This is why Jeffrey's theory has come to be known as *evidential expected utility (EEU) theory*: one might prefer an act in part because it gives one evidence that one's preferred outcome obtains. Many have argued that this feature of the theory is problematic, and the problem can be brought out by a case know as *Newcomb's problem* (first discussed by Robert Nozick (1969)).

Here is the case. You are presented with two boxes, one closed and one open so that you can see its contents; and you may choose either to take only the closed box, or to take both boxes. The open box contains $1000. The contents of the closed box were determined as follows. A predictor predicted ahead of time whether you would choose to take the one box or both; if he predicted that you would take just the closed box, he's put $1M in the closed box, but if he predicted that you would take both, he's put nothing in the closed box. Furthermore, you know that many people have faced this choice and that he's predicted correctly every time.

Assuming you prefer more money to less, E EU theory recommends that you take only one box, since the relevant conditional probabilities are one and zero (or close thereto): p(there is $1M in the closed box |you take one box) $\approx$ 1, and p(there is $0 in the closed box |you take two boxes) $\approx$ 1. But many think that this is the wrong recommendation. After all, the closed box already contains what it contains, so your choice is between receiving whatever is in that box and receiving whatever is in that box plus an extra thousand dollars. Taking two boxes *dominates* taking one box, the argument goes: it is better in every possible world. We might diagnose the mis-recommendation of EEU theory as follows: p($1M |one box) is high because taking one box is correlated with getting $1M, but taking one box cannot cause $1M to be in the box because the contents of the box have been already determined; and so EEU theory gets the recommendation wrong because conditional probability does not distinguish between correlation and causation. Not everyone accepts that two-boxing is the correct solution: those who advocate one-boxing point out that those who take only one box end up with more money, and since rationality ought to direct us to the action that will result in the outcome we prefer, it is rational to take only one box. However, those who advocate two-boxing reply that even though those who take only one box end up with more money, this is a case in which they are essentially rewarded for behaving irrationally.

For those who advocate two-boxing, one way to respond to this problem is to modify E EU theory by adding a condition like *ratifiability* (Jeffrey 1983: pp. 19–20), which says that one can only pick an act if it has the highest E EU theory on the supposition that one has chosen it. However, this does not solve the general problem of distinguishing *A*'s being evidentially correlated with *S* from *A*'s causing *S*. To yield the two-boxing recommendation in the Newcomb case, as well as to address the more general problem, Allan Gibbard and William Harper (1978/1981) proposed *causal expected utility theory*, drawing on a suggestion of Robert Stalnaker (1972/1981). Causal expected utility theory enjoins an agent to maximize:

$$u(A) = \sum_{i=1}^{\infty} p(A\Box \rightarrow S_i)\, u(S_i \,\&\, A)$$

where $p(A\Box\rightarrow S_i)$ stands for the probability of the counterfactual "If I were to do A then S_i would happen."[10] Armendt (1986) provided a representation theorem for the new theory, and Joyce (1999) provided a unified representation theorem for both evidential and causal expected utility theory.

Causal expected utility theory recommends two-boxing if the pair of counterfactuals "If I were to take one box, there would be \$0 in the closed box" and "If I were to take two boxes, there would be \$0 in the opaque box" are assigned the same probability, and similarly for the corresponding pair involving \$1M in the opaque box. This captures the idea that the contents of the closed box are independent of the agent's choices, and vindicates the reasoning that taking two boxes will result in an extra thousand dollars:

$$u(1box) = p(1box\Box\rightarrow C(\$0))\,u(\$0) + p(1box\Box\rightarrow C(\$1M))\,u(\$1M)$$

$$u(2boxes) = p(2boxes\Box\rightarrow C(\$0))\,u(\$1K) + p(2boxes\Box\rightarrow C(\$1M))\,u(\$1M+\$1K)$$

To get this result, it is important that the counterfactuals in question are what Lewis (1981) calls "causal" counterfactuals rather than "back-tracking" counterfactuals. For there are two senses in which the counterfactuals "If I were to take one box, there would be \$0 in the closed box" and "If I were to take two boxes, there would be \$0 in the closed box" can be taken. In the back-tracking sense, I would reason from the supposition that I take one box *back to* the conclusion that the predictor predicted I would take one box, and I would assign a very low credence to the former counterfactual; but I would by the same reasoning assign a very high credence to the latter. In the causal sense, I would hold fixed facts about the past, since I cannot now cause past events, and the supposition that I take one box would not touch facts about what the predictor did; and by this reasoning I would assign equal credence to both counterfactuals.

It is worth considering how Savage's original theory would treat the Newcomb problem. Savage's theory uses unconditional credences, but correctly resolving the decision problem depends on specifying the states, outcomes, and acts in such a way that states are independent of acts. So, in effect, Savage's theory is a kind of causal decision theory. Indeed, Lewis (1981: p. 13) thought of his version of causal decision theory as returning to Savage's unconditional credences, but building the correct partition of states into the formalism itself rather than relying on an extra-theoretical principle about entity-specification.

All of the modifications mentioned here leave the basic structure of the theory intact – probability and utility are multiplied and then summed – and treat both p and u as subjective, so we can put them all under the heading of *subjective expected utility theory* (hereafter EU theory).

How should we understand the two functions, p and u, involved in EU theory? In the case of the probability function, although there is debate over whether p is *defined* by preferences ("betting behavior") via a representation theorem or whether preferences are merely a way to *discover* p, it is widely acknowledged that p is supposed to represent an agent's beliefs. In the case of the utility function, there are two philosophical disagreements. First, there is a disagreement about whether the utility function is defined by or merely discovered from preferences. If one thinks the utility function is defined by preferences, there is a further question about whether it is merely a convenient way to represent preferences or

[10] Other formulations of causal expected utility theory include that of Lewis (1981) and Skyrms (1982).

whether it refers to some pre-theoretical, psychologically real entity, such as strength of desire or perceived degree of satisfaction. Functionalists, for example, hold that utility is (at least partially) constituted by its role in preferences but also hold that utility is psychologically real. Since the term "realism" is sometimes used to refer to the view that utility is independent of preferences, and sometimes used to refer to the view that utility is a psychologically real quantity, I will use the following terminology. I will call the view that utility is discovered from preferences *non-constructive realism* and the view that utility is defined from preferences *constructivism*. I will call the view that utility does correspond to something psychologically real *psychological realism* and the view that utility does not refer to any real entity *formalism*.[11] Non-constructive realist views will be psychologically realist as well; however, functionalism counts as a constructivist, psychological realist view. Hereafter, when I am speaking of psychological realist theories, I will speak as if utility corresponds to desire, just as subjective probability corresponds to belief, though there may be other proposals about what utility corresponds to.

36.3 The Norm of Normative Decision Theory

Representation theorems connect preferences conforming to a set of axioms on the one hand to utilities and probabilities such that preferences maximize expected utility on the other. Thus, representation theorems give us an equivalent way to state the prescription that one ought to maximize expected utility: one ought to have preferences that accord with the axioms. The upshot of this equivalence depends on which theory of utility one adopts. For psychological realists, both formulations of the norm may have some bite: the "maximization" norm is about how preferences ought to be related to beliefs and desires, and the "axiom" norm is an internal requirement on preferences. For formalists, since there is really no such thing as utility, the *only* sensible formulation of the norm is as the axiom norm. But for both interpretations, an important advantage of the representation theorems is that judgments about whether an agent did what she ought, as well as arguments about whether EU theory identifies a genuine prescription, can focus on the axioms.

A point of clarification about the equivalent ways to state the norm of EU theory is needed. "Maximize expected utility" admits of two readings, one narrow-scope ("Given your utility function, maximize its expectation") and one wide-scope ("Be such that there is a utility function whose expectation you maximize"). And the axiom norm is only equivalent to the wide-scope maximization norm. For the narrow-scope norm to apply in cases in which one fails to live up to it, one must be able to count as having a utility function even when one does not maximize its expectation. Clearly, this is possible according to

[11] The term constructivism comes from Dreier (1996), and the term formalism comes from Hanssen (1988). Bermúdez (2009) uses "operationalism" for what I call formalism. Zynda (2000) uses "strong realism" for what I call non-constructive realism and "weak realism" for what I call psychological realism.

the non-constructive realist.[12] I will also show that in many cases, it is possible according to all psychological realists.

We can note the historical progression: in its original formulations, decision theory was narrow-scope, and the utility function (or its analogue) non-constructive realist: money had an objective and fixed value. However, wide-scope, constructivist views are most popular nowadays. Relatedly, whereas originally a central justification of the norm was via how well someone who followed it did over the long run, such justifications have fallen out of favor and have been replaced by justification via arguments for the axioms.

One final point of clarification. So far, we have been talking about the relationship of beliefs and desires to preferences. But one might have thought that the point of a theory of decision-making was to tell individuals what to *choose*. The final piece in the history of decision theory concerns the relationship between preference and choice. In the heyday of behaviorism, Samuelson's (1938) idea of "revealed preference" was that preference can be cashed out in terms of what you would choose. However, nowadays philosophers mostly think the connection between preference and choice is not so tight. Throughout the rest of this chapter, I will use preference and choice interchangeably, while acknowledging that I take preference to be more basic and recognizing that the relationship between the two is not a settled question.

There are two ways to take the norm of normative decision theory: to guide to one's own actions or to assess from a third-person standpoint whether a decision-maker is doing what she ought.[13] Having explained the norm of normative decision theory, I now turn to the question of what sort of "ought" it is supposed to correspond to.

36.4 Rationality

Decision theory is supposed to be a theory of rationality; but what concept of rationality does it analyze? Decision theory is sometimes said to be a theory of *instrumental* rationality – of taking the means to one's ends – and sometimes said to be a theory of *consistency*. But it is far from obvious that instrumental rationality and consistency are equivalent. So it is worth spending time on what each is supposed to mean and how EU theory is supposed to analyze each; and in what sense instrumental rationality and consistency come to the same thing.

[12] However, there is an additional problem with the wide-scope norm for the non-constructive realist: maximizing the expectation of some utility function doesn't guarantee that you've maximized the expectation of your own utility function. The connection between the utility function that is the output of a representation theorem and the decision-maker's actual utility function would need to be supplemented by some principle, such as a contingent version of Christensen's (2001) "Representational Accuracy" or by his "Informed Preference."

[13] Bermúdez (2009) distinguishes these as two separate uses: what I call using normative decision theory to guide one's own actions he calls the "action-guiding" use, and what I call using normative decision theory for third-person assessment he calls the "normative" use; however, he includes more in the normative use of decision theory than just assessing whether the agent has preferences that conform to the norm of EU theory, such as assessing how well she set up the decision problem and her substantive judgments of desirability.

Let us begin with instrumental rationality and with something else that is frequently said about decision theory: that it is "Humean." Hume distinguished sharply between reason and the passions and said that reason is concerned with abstract reasoning and with cause and effect; and while a belief can be contrary to reason, a passion (or in our terminology, a desire) is an "original existence" and cannot itself be irrational. As his famous dictum goes, "Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger."[14] Hume thinks that although we cannot pass judgment on the ends an individual adopts, we can pass judgment if she chooses means insufficient for her ends. To see how decision theory might be thought to provide this kind of assessment, consider the psychological realist version of the theory in which an individual's utility function corresponds to the strengths of her desires. This way of thinking about the theory gives rise to the natural suggestion that the utility function captures the strength of an agent's desires for various ends, and the dictum to maximize expected utility formalizes the dictum to prefer (or choose) the means to one's ends.

The equivalence of preferring the means to one's ends and maximizing expected utility is not purely definitional. True, to prefer the means to one's ends is to prefer the act with the highest *utility*: to prefer the act that leads to the outcome one desires most strongly. However, in the situations we are concerned with, it is not clear which act will lead to which outcome – one only knows that an act will lead to a particular outcome if a particular state obtains – so one cannot simply pick the act that will lead to the preferred outcome. Therefore, there is a real philosophical question about what preferring the means to your ends requires in these situations. EU theory answers this substantive question by claiming that you ought to maximize the *expectation* of the utility function relative to your *subjective* probability function. So if we cash out EU theory in the means-ends idiom, it requires you to prefer the means that will, *on average and by your own lights*, lead to your ends. It also requires that you *have* a consistent subjective probability function and that the structure of your desires is such that a number can be assigned to each outcome. So it makes demands on three kinds of entities: beliefs, desires, and preferences given these. This formulation of the maximization norm is compatible with both the narrow-scope and the wide-scope reading: if in concert with Hume's position we think that desires cannot be changed by reason, there will be only one way to fulfill this requirement; but if we think that the agent might decide to alter her desires, there will be multiple ways to fulfill this requirement.

A more contemporary formulation of the idea that decision theory precisifies what it is to take the means to one's ends is that decision theory is *consequentialist*. This is to say that decision theory holds that acts must be valued only by their consequences. An important justification of EU maximization as the unique consequentialist norm, and a justification that formalists and psychological realists can both avail themselves of, comes from Hammond (1988). Hammond considers sequential decision problems (decision problems in "extensive" form rather than "normal" form), where decision-makers are not choosing only once but instead can revise their plan of action as new information comes in. He argues that the assumption that decision-makers value acts only for their consequences,

[14] Hume (1731/1978: p. 416).

when cashed out in terms of plausible principles about sequential choice, entails the axioms of EU theory.[15]

Even in the case of choice at a time, we can think of the axioms as representing an attempt to formalize necessary conditions for preferring the means to one's ends. I don't have space to pursue this thought in detail, but here is one example of what I have in mind. Consider the requirement of state-wise dominance, which says roughly that if act f is weakly preferred to act g in every state, and strictly preferred in some state that has positive probability, then you ought to strictly prefer f to g (this is a necessary condition of being representable as an EU maximizer). One plausible way to state what's wrong with someone whose preferences don't conform to this requirement is that she fails to prefer what she believes is superior in terms of satisfying her preferences, or she fails to connect her preferences about means to her preferences about ends. Not all of the axioms can be straightforwardly argued for in this way, but this can be a helpful way to think about the relationship of the axioms to instrumental rationality.

Thus, normative EU theory may be supported by arguments to the effect that the maximization norm or the axiom norm explicates instrumental rationality (leaving aside whether these arguments are ultimately successful). The other notion of rationality that decision theory is often described as analyzing is *consistency*, and the axiom formulation of the norm coheres well with this notion. To understand why, it is helpful to consider the related idea that the principles of formal logic analyze what it is to have consistent binary beliefs. There are two important standards at work in binary belief. First, an agent ought to believe what is reasonable to believe, given her evidence. This is a requirement about the substance of her beliefs, or about the content of her beliefs vis-à-vis her evidence or what the world is like. Secondly, an agent's beliefs ought to be consistent with one another in the sense elucidated by logic. This is a requirement about the structure of her beliefs, or about the content of her beliefs vis-à-vis the content of her other beliefs. This isn't to say that agents must be logically perfect or omniscient, or that there is an external standard of adherence to the evidence, but the point is that these are two different kinds of norms and we can separately ask the questions of whether a believer conforms to each.

Similarly, in evaluating preferences over acts, there are two questions we might ask: whether an agent's preferences are reasonable, and whether they are consistent. Here, the axioms of decision theory are supposed to play a parallel role to that played in beliefs by the axioms of logic: without regard to the content of an agent's preferences, we can tell whether they obey the axioms. Just as the axioms of logic are supposed to spell out what it is to have consistent binary beliefs, so too are the axioms of decision theory supposed to spell out what it is to have consistent preferences.

There are several ways in which it might be argued that the axioms correctly spell out what it is to have consistent preferences.[16] One classic argument purports to show that violating one or more of them implies that you will be the subject of a "money pump," a situation in which you will find a series or set of trades favorable but will disprefer the entire package,

[15] For further discussion of this type of argument, see Seidenfeld (1988), McClennen (1990), and Levi (1991).

[16] Not all philosophers think that *arguing* for this conclusion is the right way to proceed. For example, Patrick Maher (1993: pp. 62, 83) suggests that no knock-down intuitive argument can be given in favor of EU theory, but that we can justify it by the fruits it produces.

usually because taking all of them results in sure monetary loss for you.[17] This amounts to valuing the same thing differently under different descriptions – as individual trades on the one hand and as a package on the other – and is thought to be an internal defect rather than a practical liability.[18] A different argument, due to Peter Wakker (1988), purports to show that violating one of the axioms will entail that you will avoid certain types of cost-free information.

I propose that consistency in preferences is an amalgamation of consistency in three different kinds of entities: consistency in preferences over outcomes, consistency in preferences about which event to bet on, and consistency in the relationship between these two kinds of preferences and preferences over acts.[19] Or, psychological realists might say: consistency in desires, consistency in beliefs, and consistency in connecting these two things to preferences. Aside from the fact that adhering to the axioms does produce three separate functions (a utility function of outcomes, a probability function of states, and an expectational utility function of acts), which is not decisive, I offer two considerations in favor of this proposal. First, arguments for each of the axioms can focus more or less on each of these kinds of consistency. For example, an argument that transitivity is a rational requirement doesn't need to say anything about beliefs or probability functions. Secondly, a weaker set of axioms than those of EU theory will produce a consistent probability function without a utility function relative to which the agent maximizes EU; and a weaker set of axioms than those of EU theory will produce a utility function of outcomes without a probability function relative to which the agent maximizes EU; and a weaker set of axioms than those of EU theory will produce a utility function and a probability function relative to which an agent maximizes something other than EU.[20] Therefore, even if the justifications for each of the axioms are not separable into those based on each of the three kinds of consistency, the kinds of consistency are formally separable. And here is a difference, then, between logic and decision theory: logical consistency is an irreducible notion, whereas decision-theoretic consistency is a matter of being consistent in three different ways.[21]

Here, then, are the ways in which instrumental rationality and consistency are related. First, and most obviously, there are arguments that each is analyzed by EU theory; if these arguments are correct, then instrumental rationality and consistency come to the same thing. Secondly, given that consistency appears to involve consistency in the three kinds of

[17] Original versions of this argument are due to Ramsey (1926/1931) and de Finetti (1937/1980).

[18] See Christensen (1991), although he is mostly concerned with this type of argument as it relates to the subjective probability function.

[19] By a preference to bet on one event rather than another, I mean a preference to receive a favored outcome on the former rather than to receive that outcome on the latter.

[20] For an axiomatization of a theory that yields a probability function for a certain kind of non-EU maximizer, see Machina and Schmeidler (1992). For an axiomatization of a theory that yields a utility function for an agent who lacks a subjective (additive) probability function, see Gilboa (1987), or any non-expected utility theory that uses subjective decision weights that do not necessarily constitute a probability function. For an axiomatization of a theory that yields a utility and probability function relative to which an agent maximizes a different functional, see Buchak (2013).

[21] Note, however, that for non-constructive realists, there could be a case in which two of these things are inconsistent in the right way but preferences are still consistent. See Zynda (2000: pp. 51–60), who provides an example of an agent whose beliefs are not governed by the probability calculus and whose norm is not expected utility maximization relative to his beliefs, but who has the same preferences as someone who maximizes EU relative to a probability function.

entities instrumental rationality is concerned with, consistency in preferences can be seen as an internal check on whether one really prefers the means to one's ends relative to a set of consistent beliefs and desires.[22]

It was noted that even if binary beliefs are consistent, we might ask the further question of whether they are reasonable. Can a similar question be applied to preferences? Given that consistency applies to three entities, the question of reasonableness can also be separated into three questions: whether the subjective probability function is reasonable, whether the utility function is reasonable, and whether one's norm is reasonable. The reasonableness question for subjective probability is an analogue of that for binary beliefs: are you in fact proportioning your beliefs to the evidence? For the formalist, the reasonableness question for utility, if it makes sense at all, will really be about preferences. But for the psychological realist, the reasonableness question for utility might be asked in different ways: whether the strengths of your desires in fact track what would satisfy you, or whether they in fact track the good. In EU theory, there is only one norm consistent with taking the means to your ends – maximize expected utility – so the reasonableness question appears irrelevant; however, with the introduction of alternatives to EU theory, we might pose the question, as I will discuss in section eight.

36.5 Interpretive Decision Theory

The major historical developments in normative decision theory mostly came from considering the question of what we ought to do. By contrast, another strand of decision theory was moved forward by philosophical questions about mental states and their relationship to action.

In 1926, Frank Ramsey was interested in a precise way to measure degrees of belief, since the prevailing view was that degrees of belief weren't appropriate candidates to use in a philosophical theory unless there was a way to measure them in terms of behavior. Ramsey noted that since degrees of belief are the bases of action, we can measure the degree of a belief by the extent to which the individual would act on the belief in hypothetical circumstances. Ramsey created a method whereby a subject's preferences in hypothetical choice situations are elicited and her degrees of belief (subjective probabilities) are inferred through these, without knowing her values ahead of time. For example, suppose a subject prefers getting a particular prize to not getting that prize, and suppose she is neutral about seeing the heads side of a coin or the tails side of a coin. Then if she is indifferent between the gamble on

[22] Compare to Niko Kolodny's proposal that wide-scope requirements of formal coherence as such may be reducible to narrow-scope requirements of reason. The "error theory" in Kolodny (2007) proposes that inconsistency in beliefs reveals that one is not adopting, on some proposition, the belief that reason requires; and the error theory in Kolodny (2008) proposes that inconsistency in intentions reveals that one is not adopting the intention that reason requires. Direct application of Kolodny's proposal to the discussion here is complicated by the fact that some see the maximization norm as wide-scope and some as narrow-scope. But those who see it as narrow-scope may take a Kolodny-inspired line and hold that consistency in preferences is merely an epiphenomenon of preferring that which you have reason to prefer, given your beliefs and desires.

which she receives the prize if the coin lands heads and the gamble on which she receives the prize if the coin lands tails, it can be inferred that she believes to equal degree that the coin will land heads as that it will land tails, i.e., she believes each to degree 0.5. If she prefers getting the prize on the heads side, it can be inferred that she assigns a greater degree of belief to heads than to tails.

Generalizing the insight that both beliefs and values can be elicited through preferences, Ramsey presented a representation theorem. Ramsey's theorem was a precursor to Savage's, and like Savage's theorem, Ramsey's connects preferences to a probability function and a value function, both subjective. Thus, like the normative decision theorists that came after him, Ramsey saw that maximizing expected utility with respect to one's personal probability and utility functions is equivalent to having preferences that conform to certain structural requirements. However, Ramsey was not interested in using the equivalence to reformulate the maximization norm as a norm about preferences. Rather, he *assumed* that preferences *do* conform to the axioms, and used the equivalence to discover facts about the agent's beliefs and desires.

Related to Ramsey's question of how to measure beliefs is the more general question of how to attribute mental states to individuals on the basis of their actions. Donald Davidson (1973) coined the term "radical interpretation" (a play on W.V.O. Quine's "radical translation") to refer to the process of interpreting a speaker's beliefs, desires, and meanings from her behavior. For Davidson, this process is constrained by certain rules, among them a principle about the relationship between beliefs and desires on the one hand and actions on the other, which, as David Lewis (1974) made explicit, can be formalized using expected utility theory. Lewis's formulation of the "Rationalization Principle" is precisely that rational agents act so as to maximize their expectation given their beliefs and desires. Thus, Ramsey's insight became a part of a more general theory about interpreting others. For theorists who make use of EU theory to interpret agents, maximizing EU is constitutive of (rational) action; indeed, Lewis (1974: p. 335) claims that the Rationalization Principle has a status akin to analyticity.

An immediate tension arises between the following three facts. First, for interpretive theorists, anyone who cannot be interpreted via the Rationalization Principle will count as unintelligible. Secondly, actual human beings are supposed to be intelligible; after all, the point of the project is to formalize how we make sense of another person. Thirdly, actual human beings appear to violate EU theory; otherwise, the normative theory wouldn't identify an interesting norm.

One line to take here is to retain the assumption that it is analytic that agents maximize EU, and to explain away the apparent violations. We will see a strategy for doing this in the next section, but I will argue there that adopting this strategy in such a way as to imply that EU maximization cannot be violated leads to uninformative "interpretations." A more promising line starts from the observation that when we try to make sense of another person's preferences, we are trying to make sense of them as a whole, not of each considered in isolation. Consider an agent whose preferences mostly conform to the theory but fail to in a few particular instances, for example, an individual who generally gets up at 7 a.m. to go for a run but occasionally oversleeps her alarm. We would say that she prefers to exercise in the morning. Or consider an individual who generally carries an umbrella when the chance of rain is reported as at least 50%, but who on one occasion leaves it at home when she thinks it is almost certain to rain. We would say that she considers the burden of carrying

around an umbrella only moderate in comparison to how much she does not like to get wet. In general, if a large set of an individual's preferences cohere, the natural thing to say is that she has the beliefs and desires expressed by those preferences but that her preferences occasionally fail to match up with her beliefs and desires, perhaps because she is on occasion careless or confused or weak of will.

This suggests what interpretive theorists ought to do in the case of non-ideal agents: take an agent's actual preferences, consider the closest "ideal" set of preferences – the closest set of preferences that do conform to the axioms – and infer the agent's beliefs and desires from these. Thus the theorist will interpret the agent as being as close to ideally rational as possible: we might say, as maximizing expected utility in general, but as occasionally failing to do so. Furthermore, this allows us to interpret the agent as failing to maximize the expectation of her utility function on occasion – that is, as *having* desires on this occasion but failing to prefer in accordance with them – precisely because her obeying the axioms in a large set of her preferences or having a closest ideal counterpart points to a utility function that is genuinely hers. I note that "closest" here might be cashed out either as differing least from the agent's actual preferences, or as preserving the values that the agent would endorse in a clear-headed frame of mind, or as best according with other facts about her psychology, such as her utterances. I also note that in some cases there will be no close counterpart, and it will be precisely these cases in which the interpretive theorist will count the agent as unintelligible, as not intentionally acting.

There is one problem with this method, however. It does not allow us to interpret an agent as having genuinely inconsistent beliefs or desires, only as failing to have preferences that accord with them on occasion. While I don't have space to fully explore the possibilities here, there seem to me to be several options. First, and perhaps less plausibly, an interpretive theorist might postulate that an individual "really" has a coherent set of beliefs and desires, though these aren't always correctly translated into preferences. Secondly, one might postulate that an agent's degree of belief in a proposition is derived from (the closest ideal set to) some privileged set of preferences; for example, as many propose, that $p(E)$ is derived from the bets in small amounts of money one is willing to make on E. And similarly, perhaps, for desire, although it is harder to say what the privileged set might be. Finally, if some of one's preferences cluster towards one ideal counterpart and some towards another, along a natural division, and we could postulate that the agent is of two minds in a very particular way.

Decision theory appears in philosophy in two different strands. The normative theorist is interested in what your preferences ought to be given your beliefs and desires or given other of your preferences. Adopting EU maximization or conformity to the axioms as the correct norm, she says that you ought to prefer that which maximizes expected utility, and she is interested in which acts would do so; or she says that you ought to have consistent preferences, and she is interested in which sets of preferences are consistent. The interpretive theorist is interested in discovering what your beliefs and desires are from your preferences. Adopting EU maximization or conformity to the axioms as the correct principle of interpretation, she says that you do (approximately) maximize expected utility or have consistent preferences, and she is interested in what beliefs and desires make it the case that you do so.

When thus described, we can see that rationality plays a different role in each use of the theory: the interpretive theorist takes it as an *assumption* that individuals are rational in the decision-theoretic sense, and the normative theorist takes decision theory as a way

to *answer* the question of whether individuals are rational. Although on the face of it this makes it seem that the two uses are in tension, I have proposed that on the best way to make sense of the interpretive project, the concept of rationality that is meant to be analyzed by EU theory is importantly different in the two projects. Specifically, the rationality norm of the normative project is "strong" in that normative theorists are interested in whether *all* of the individual's preferences adhere to it, and the rationality assumption in the interpretive project is "weak" in that interpretive theorists make the assumption that an agent more or less follows it but not the stronger assumption that she follows it exactly and always.

36.6 Outcome Descriptions

Much has been made so far of the fact that by connecting preferences to subjective utility and probability functions, we can discover how much an agent values outcomes and how likely she takes various states to be. But one issue that has not yet been remarked upon is that just as how the agent values the outcomes and views the world are not intrinsic features of any situation she faces, neither is how the agent conceptualizes the outcomes.

An illustrative example is due to John Broome (1991: pp. 100–1). Maurice is offered choices between various activities. If the choice is between going sightseeing in Rome and going mountaineering, Maurice prefers to go sightseeing, because mountaineering frightens him. If the choice is between staying home and going sightseeing, he prefers to stay home, because Rome bores him. However, if the choice is between mountaineering and staying home, he prefers to go mountaineering, because he doesn't want to be cowardly.

If we consider Maurice's preferences among Rome, home, and mountaineering, they appear to be intransitive: he prefers Rome to mountaineering, home to Rome, and mountaineering to home. Given that transitivity is necessary for EU maximization, the interpretive theorist is unable to make sense of him given the preferences as stated; but his motivation is perfectly comprehensible (we've just described it). In addition, the normative theorist must automatically count Maurice's preferences as irrational, without considering whether his reasons for them make sense; but it is not clear – at least not without further argument – that there really is anything wrong with his preferences.

Here is what each theorist ought to be able to say: for Maurice, choosing home when the alternative is going to Rome is different from choosing home when the alternative is mountaineering. Therefore, there are really (at least) four options involved in this decision problem: Rome, home-when-the-alternative-is-Rome, home-when-the-alternative-is-mountaineering, and mountaineering. And Maurice's preferences among these options are not, as far as we know, intransitive. The lesson is that insofar as we are concerned with capturing the agent's actual beliefs and desires, we cannot assume that there is a privileged description of outcomes independent of how the agent himself sees them. Furthermore, insofar as we are interested in determining whether an agent is genuinely consistent or genuinely prefers the means to his ends, we cannot automatically rule out his caring about certain features of outcomes. What the agent believes and desires is what we are trying to determine, and that includes what the agent believes about the choices he faces.

Thus, there is an additional "moving piece" in the interpretation of an agent or in a judgment about whether his preferences are rational: how he sees the outcomes. This poses two related challenges. The first is about how to settle on the correct interpretation of the agent's preferences. The second has to do with the extent to which individuating outcomes more finely commits the agent to having preferences in choice situations that could never even in principle be realized, and how we ought to treat these preferences. I will discuss these issues in reverse order.

To illustrate the second problem, notice that it is assumed in decision theory that preferences are complete: for any two options, a decision-maker must prefer one to the other or be indifferent. This means that if "home-when-the-alternative-is-Rome" and "home-when-the-alternative-is-mountaineering" are to count as options in some of the choice problems described above, the decision-maker must prefer one to the other or be indifferent. But one could never actually face a choice between these two options, by definition. Broome (1991) refers to preferences of this sort as "non-practical preferences." I will not discuss the metaphysics of these preferences: although there are interesting questions here, they do not obviously bear on the issues this chapter has been focusing on. But the epistemology of these preferences is important, because it will make a difference to how we resolve the interpretive problem more generally. There will be a divide between those who think that which of these options an agent prefers is up to the agent, and those who think that which of these options an agent prefers is up to the decision theorist to fill in; and where one falls in respect of this divide will determine how much freedom a decision theorist has to interpret an agent's preferences.

The other problem, then, is how to settle on an interpretation of the agent's preferences. As we've just seen, we cannot dictate that the theorist's initial presentation of the outcomes is how the agent sees them. However, if we allow that the agent makes maximally fine distinctions between outcomes, then none of the outcomes will be the subject of more than one practical preference. For example, choosing each outcome in a pair-wise choice always involves rejecting the other alternative. If the agent's non-practical preferences are up to the theorist to fill in, then this will mean that the agent can never fail to maximize expected utility nor fail to satisfy the axioms, since no practical preferences will be inconsistent with each other.

If the norm of EU theory were impossible to violate, the normative theory would lose its bite, since it will be trivially true that every agent adheres to the norm. But would this also be a problem for the interpretive EU theorist? Some might say that it wouldn't be; indeed, that EU maximization is trivially satisfied would lend support to the idea that it is a good interpretive assumption that agents *actually* maximize EU. But there are at least two problems with this approach for the interpretive theorist. The first is that we will be unable to tell the difference between when an individual is trying to maximize EU (or follow the axioms) but making a mistake and when she is aiming at something else,[23] although perhaps this is okay if it is argued that to act at all is to maximize EU. The second problem is that allowing outcomes to be individuated maximally finely means that "deriving" an agent's beliefs and desires from her preferences won't be very informative. Her practical preferences in combination with each possible filling out of her non-practical preferences will give rise

[23] See Hurley (1989: pp. 55–83)

to a unique (up to positive affine transformation) utility and probability function, by the representation theorems. But there may be many possible fillings out. Therefore, there will be multiple and incompatible ways to interpret her beliefs and desires. And on the level of preferences, knowing what she prefers in one particular context won't tell us anything about what she prefers in an only slightly different context, so we won't get a very robust explanation of her psychology. In either case, the theory is rendered uninformative: we cannot make much sense of what the agent is doing.

Most philosophers accept that either the theorist's ability to individuate outcomes or the theorist's ability to set non-practical preferences must be constrained. To constrain them, one can either introduce a rule about when two outcomes are allowed to count as different, or allow that outcomes can be individuated as finely as possible but introduce a rule about what non-practical preferences the theorist can interpret the agent as having. For most purposes, these come to the same thing, since refusing to allow that x and y are different outcomes and requiring that the correct interpretation of the agent makes her indifferent regarding the choice between x and y permit the same sets of practical preferences. But there are two very different types of constraining rules that the theorist could introduce (this distinction crosscuts the distinction just mentioned). To see this, consider the following suggested rules:

R1: Outcomes should be distinguished as different if and only if they differ in a way that makes it rational to have a preference between them (Broome 1991: p. 103).
R2: Outcomes should be distinguished as different if and only if the agent actually has a preference between them (Dreier 1996: p. 260).
R3: Outcomes should be distinguished as different if and only if they differ in regard to properties that are desired or undesired by the agent (Pettit 1991: p. 165).

Maurice's preferences can be accommodated by EU theory according to rule R1 only if it is rational for Maurice to care about what option he turns down when he decides to stay at home, according to rule R2 only if he in fact does care about what option he turns down when he decides to stay at home, and according to rule R3 only if turning down some option instantiates a property he in fact cares about.

Rules R2 and R3 make the possibility of distinguishing outcomes dependent on the agent's internal state, whereas R1 makes this possibility dependent on some objective feature of the agent's situation. Rules such as R1 that introduce an external constraint on interpretation might be seen as principles of charity for interpretation: we should interpret an agent as making a distinction only if it is rational to make that distinction. Since these "externalist" rules restrict preferences beyond what the agent herself values, Broome has rightly pointed out that they are against the spirit of Humeanism (Broome 1993). Of course, rules like R2 and R3 can be applied only if the theorist has access to the agent's non-practical preferences or other relevant properties of her internal state. Therefore, using these "internalist" rules relies on the theorist knowing more about an agent's psychology than the externalist rules do.

The same strategy can be applied to ensuring that the norm of normative decision theory is not trivial. As long as there is a restriction on when two outcomes can count as different, there will be sets of preferences that violate the norm of EU theory. Which type of rule to adopt will depend on the use to which normative decision theory is being put: if the

theorist is using it to assess an agent, whether the theorist can rely on an internalist rule will depend on how much she can know about the agent's internal state, and if the agent is using normative decision theory to guide her own actions, whether she can rely on an internalist rule will depend on how much introspective access she has to her own internal state.

36.7 Descriptive Decision Theory

Although a third type of decision theory, descriptive decision theory (which sometimes goes by the name "behavioral economics"), is largely the provenance of psychology and economics rather than philosophy, it is important to say something about it both for completeness and to make clear the contrast with interpretive decision theory.

Like interpretive decision theory, descriptive decision theory is interested in describing the behavior of individuals rather than in what they ought to do. However, there is an important difference between the two approaches, which can be seen in how they have responded to findings that actual agents fail in reliable ways to maximize expected utility. Whereas interpretive decision theory has retained EU maximization as the guiding principle of interpretation, and in many cases pushed for a more complex interpretation of outcomes (as described in the previous section), descriptive decision theory has by and large abandoned expected utility maximization as an unrealistic assumption about agents and has proposed alternatives.

I do not have space to go into the alternatives to EU theory that descriptive theorists have proposed (see Sugden (2004) and Schmidt (2004) for helpful surveys), but it is worth saying two ways in which these alternatives tend to differ from EU theory. First, while they generally include a function that plays the role of utility and a function that plays the role of probability, they either subject these to different constraints (e.g. the "probability" function needn't be additive) or else combine them in a non-expectational way. Secondly, at least one notable alternative, Kahneman and Tversky's (1979) *prospect theory*, posits an "editing phase" during which the decision-maker simplifies the alternatives using various heuristics before subjecting them to the maximization schema.

The differing responses of the two types of theorists to purported violation reveals two important differences between the aims of descriptive decision theory and the aims of interpretive decision theory. First, descriptive theorists are generally interested in building parsimonious models of preferences, and they are less concerned than interpretive theorists with interpreting the utility and probability functions as desires and beliefs. Interpretive theorists, by contrast, are primarily interested in extracting desires and beliefs with few or no initial assumptions, including assumptions about how an agent views the outcomes; and in doing so need be only as parsimonious about outcomes as an agent's actual psychology is. Therefore, descriptive decision theorists are more inclined to treat the outcomes (for them, generally, monetary values) as theoretical bedrock, and interpretive decision theorists are more inclined to treat the rationalization principle as theoretical bedrock. It is worth noting that for the same reasons that economists concerned merely with modeling behavior will be uninterested in the interpretive project, formalists will also not be interested in the interpretive project, since for them, there aren't any interesting entities worth discovering.

The other difference, which I will discuss further in the next section, concerns predictable deviation from rationality. Roughly, if agents predictably have preferences against the dictates of rationality, the descriptive theorist will want to include this as part of her model, since it is an accurate characterization of what the agent does, but the interpretive theorist will not, since, recalling the discussion in section four, those preferences do not accurately reflect her beliefs and desires (though predictable deviations may be included somewhere in the theory of action). Interpretive theorists are interested in characterizing the preferences of an idealized version of the agent, and descriptive theorists those of the actual, non-ideally-rational agent.

We might put these two points succinctly, although this is certainly too coarse: descriptive theorists are concerned with *prediction*, and interpretive theorists are concerned with *explanation* in terms of beliefs and desires and with discovering something about an agent's mental states.[24]

How does the descriptive project bear on the interpretive project? If the analogues of u and p in the descriptive project should be taken in a formalist vein, then the descriptive project does not have a clear bearing on the interpretive project. But insofar as the entities involved in the descriptive project can be thought of as beliefs and desires, rather than convenient ways to represent preferences, I think there is a way in which the descriptive project aids the interpretive project, and another in which it cuts against it. On the one hand, the descriptive project can help illuminate the relationship between an agent's actual choices and desires and those of her ideal counterpart. For example, one of the findings of prospect theory (Kahneman and Tversky 1979: p. 273) is that when a new frame of reference is experimentally induced, e.g., the agent believes she will receive $2000, her preferences over total amounts of money are altered in the sense that receiving less than (e.g.) $2000 will be treated as a "loss." If what this shows is that inducing a reference point causes people to underestimate their actual (subjective) utility below the reference point, then we can expect that the ideal counterpart will assign higher utility below the reference point than the actual agent in the grip of framing effects. On the other hand, if what the descriptive project reveals is that an agent cannot be interpreted as having stable beliefs and desires – beliefs and desires that are independent of the ways in which choices are presented – then the descriptive project undermines the interpretive project.

36.8 The Mutual Dependence of the Normative and Interpretive Project

In this section, I will explain how the normative and interpretive projects depend on each other. Recall that the rationality assumption in interpretive decision theory is that agents are approximately expected utility maximizers; and an agent's beliefs and desires are the p and u extracted from the preferences of her ideal counterpart. But why should we think that the beliefs and desires of an agent's ideal counterpart are *her* beliefs and desires? After

[24] Bermúdez (2009) thinks that explanation and prediction ought to be considered a single dimension of decision theory, and thus that the same formal theory must play both roles.

all, the *preferences* of her ideal counterpart aren't her actual preferences. The crucial idea is that acting consists not in actually taking the means to your ends, but in *aiming at* doing so. Therefore, the preferences of an agent's ideal counterpart are the preferences that she ought to be thought of as aiming at satisfying when she acts. This doesn't mean that she consciously aims at satisfying these preferences, or even that she consciously takes herself to be maximizing expected utility; rather, she participates in an activity (acting) which is constituted by aiming at being an EU maximizer. In the means-ends idiom, to act is to aim to take the means to your ends (or more precisely the means that will on average and by your own lights lead to your ends), even though you might sometimes fail to do so. That aiming is not the same as succeeding explains the fact that the rationality assumption in interpretive theory is not that agents are perfectly rational but rather that they are approximately rational.[25]

Now that it is clear what the rationality assumption in interpretive decision theory amounts to, it should also be clear how the interpretive project depends on the normative project. Interpretive EU theory rests on two claims. First, on the claim that action aims at conforming to the norm that analyzes what it is to take the means to one's ends or to be consistent. Secondly, on the claim that this norm is captured by EU theory, either in the maximization formulation or the axiom formulation. If we were to conclude that a different norm holds of rational preferences, then interpretive decision theory would have to follow suit in adopting that norm as the one action aims at. The interpretive project depends on the correctness of the normative theory's norm, i.e., on the normative theorist being correct about which sets of preferences are candidates for those of an agent's ideal counterpart. (Note that this is another difference between interpretive and descriptive decision theory: the latter is not at all governed by what the correct norm is.)

The normative project, if it is able to say anything interesting about agents who fall short of the norm, also depends on the interpretive project. This is because identifying how an agent is falling short of the norm depends on correctly interpreting what her beliefs and desires are. Clearly, a simple "yes or no" answer to the question of whether the agent is doing what she ought doesn't rely on discovering the agent's beliefs and desires: if her preferences don't conform to the axioms, then she fails to do what she ought. The formalist will say that normative decision theory ends here. However, there are two additional questions the psychological realist might be interested in. First, what is the source of the agent's irrationality? And secondly, where should the agent go from here?

Recall that preference inconsistency could come from one (or more) of three sources: inconsistency in beliefs, inconsistency in desires, and inconsistency in the norm connecting beliefs and desires. But we need to be able understand the agent as having beliefs and desires even when she is inconsistent if we want to claim that her beliefs or her desires are inconsistent. And so if we adopt the interpretive idea that an agent's beliefs and desires can be discovered even when she is not fully in accord with the axioms by working backwards

[25] We should untangle the question of whether postulating that agents aim at EU maximization allows the preferences of an agent's ideal counterpart to reveal her beliefs and desires from the more general question of whether postulating that agents aim at whatever the correct norm of rationality is allows this. Meacham and Weisberg (2011) argue against the former claim on empirical grounds, but we might still uphold the latter claim by arguing that rationality is best analyzed by a different norm which people in fact come closer to adhering to.

from the preferences of her ideal counterpart(s), we can say in what respect she is falling short of the normative ideal. Unless one can introspect one's beliefs and desires to a precise degree, diagnosing where the irrationality comes from depends on the possibility of interpreting agents as having beliefs and desires even when they are not obeying the axioms.

Furthermore, since the interpretive use of the theory allows us to discover an agent's beliefs and desires, it allows us to say what sort of underlying change moving from the agent's actual preference to a set of preferences that conform to the theory involves. For example, consider an individual who is willing to add $200 to the purchase price of a car she is buying if it has a radio but would not pay $200 to have a radio installed if the car came without one at the cheaper price.[26] Let us assume the closest ideal agent prefers $200 to the radio, so that the actual agent desires $200 more than she desires the radio. The narrow-scope norm says she ought to alter her preference concerning the initial purchase. The wide-scope norm of decision theory is more permissive. It says that she can resolve the irrationality by adopting any consistent set of preferences: so she can alter her preference concerning the initial purchase and retain the rest of her preferences, or she can keep that preference and alter the rest of her preferences. But even if we adopt the wide-scope norm, interpreting the agent is crucial because it allows us to say what each resolution involves: the former resolution involves conforming her preferences over acts to her underlying desires; the latter involves bringing her underlying desires in line with the preference about purchasing a radio with her car. This doesn't by itself show how she ought to resolve the decision, since in principle it may be that the one preference is more important than her desires, but it does tie different ways of resolving the decision to preserving specific different features of her situation.

In sum, if the normative theorist wants to say more than that the agent is not doing what she ought, or that she ought to bring her preferences in line with the axioms somehow or other but with no guidance on what considerations are involved in potential resolutions, she will have to interpret the agent.

The assumption of rationality in interpretive decision theory is that the agent aims at maximizing EU, and so approximates an EU maximizer. And the goal of rationality in normative decision theory is that the agent maximizes EU in every instance. This, then, is how the two projects are mutually dependent: that agents are approximately EU maximizers depends on EU maximization being the aim of rational action, and that agents bring their preferences into line with EU maximization in a way that is governed by reasons depends on locating the source of their current deviation, which depends on understanding what their beliefs and desires are.

Descriptive decision theory also bears on these projects. As I indicated in section seven, one thing that descriptive decision theory could reveal is that it would be seriously misguided to think of action as aiming at maximizing expected utility. This would undermine interpretive EU theory. But what would it say about rational action more generally? As mentioned, interpretive decision theory operates with two assumptions, one that action aims at adhering to the norm of rationality and one about what the norm is. If action doesn't aim at the maximization of EU, then we must drop either the assumption that action aims at the norm or the assumption that EU is the correct norm. If we keep

[26] Example adapted from Savage (1954/1974: p. 103).

the latter assumption, then it may be possible to use a descriptive theory to extract beliefs and desires and a normative theory that takes these as the real beliefs and desires and enjoins you to maximize expected utility.[27] On the other hand, we might keep the former assumption and propose a different norm, one that coheres closely enough with actual behavior that interpretive decision theory can use the norm to backwards engineer beliefs and desires despite some deviations that the correct descriptive theory predicts. Which of these positions to take will not be determined by empirical findings but by arguments about what the correct norm is, although the knowledge that humans diverge wildly from EU theory might give us reason to examine more closely whether EU is the correct norm, given how successful human behavior is in general.[28]

36.9 Challenges and Extensions

Philosophers have challenged the idea that EU theory is the correct theory of rationality. Recall that in the expected utility equation, the utility value of each outcome is weighted by the probability the decision-maker assigns to the state in which she receives that outcome, and this probability is supposed to reflect her belief about that state. This assumes two things: first, that the norm relating beliefs and desires to preferences is indeed that of EU maximization, in which the average value of a gamble suffices to determine its position in the agent's preference ranking (whether this is derived from the axioms or not); secondly, that rational beliefs are "sharp" in that they can be measured by point-probabilities. Challenges to each of these assumptions have been around for at least 50 years, but have resurfaced recently. One might similarly challenge that the structure of desire is that posited by EU theory, though I don't have space to discuss such a challenge here. Each of these challenges can be posed directly about the functions p, u, or the maximization norm, but each can also take the form of criticizing one or more of the axioms.

The first challenge is to the idea that we ought to care only about the expectation of utility, and not other "global" features of a gamble, such as its minimum utility value, its maximum utility value, or the spread or variance of utility. If utility is derived or discovered via a representation theorem, this point must take the form of or be accompanied by a challenge to one or more of the axioms of EU theory. Maurice Allais (1953/1979), taking what appears to be a non-constructive realist view of the utility function, argued that agents might care not just about the mean utility value of a gamble, but also about its variance and skewness. But his famous counterexample to EU theory (which has since become known as the Allais Paradox) poses a challenge even for constructivists, since it shows that most decision-makers violate one of the axioms of EU theory.[29] I have recently defended axioms that give rise to other maximization norms than that of EU theory (Buchak 2013): my view is that EU maximization is one of a more general class of norms, any of which a rational agent may adopt. In the same spirit as the idea that agents have subjective u and p functions,

[27] Bermúdez (2009: pp. 165–7) considers but rejects a possibility like this in his discussion of whether decision theory could play multiple roles at once.

[28] For an argument that humans did not evolve to be EU maximizers, see Okasha (2007).

[29] At least under the assumption that outcomes cannot be individuated more finely.

I propose that the norm that connects these to preferences is also up to the agent. Just as *u* and *p* are subject to structural constraints, so too is the norm; and, furthermore, the question mentioned in section four about whether a particular norm is reasonable, in addition to rational, can be posed.

The second challenge is to the idea that we ought to be "probabilistically sophisticated" in the sense of assigning to every event a point probability (or acting as if we do). Daniel Ellsberg (1961) proposed a pair of choice problems that have become known as the Ellsberg Paradox, purporting to show that when individuals lack precise information about objective probabilities, they don't act as if they make choices based on a single probability function. In recent years, the challenge has come from the side of epistemology rather than observed decision-making behavior, the idea being that our evidence is often imprecise or incomplete, so requiring precise degrees of belief would mean requiring degrees of belief that outrun the evidence.[30] Denying that we have sharp degrees of belief and that we need them in order to make rational decisions requires stating both what non-sharp (or "imprecise") degrees of belief are and how to make decisions with them.[31]

36.10 Conclusion: Decision Theor*ies*

Given both the historical progression and the issues that are currently under discussion, we ought not think of decision theory as a single theory, but rather as a collection of theories that each contain both a structural and interpretive element. The structural element describes the internal structure of several functions and the formal relationship between these functions on the one hand and preferences on the other; a formal relationship which holds just in case a particular set of axioms is satisfied. This internal structure and relationship are argued to be those that hold for rational agents. In EU theory, the posited functions are a numerically valued utility function and a point-probability function that obeys the probability calculus; and the posited relationship is that of EU maximization. The interpretive element concerns how these functions are to be interpreted: as psychologically real and in principle separate from preferences; as psychologically real and tightly connected to preferences; or merely as a representation of preferences. Whichever combination of structural element and interpretation we adopt, the underlying issues discussed in the previous few sections – the relationship between decision theory and rationality, how to individuate outcomes, and the relationship between normative decision theory and interpretive decision theory – remain the same.

References

Allais, M. (1953/1979) The Foundations of a Positive Theory of Choice involving Risk and a Criticism of the Postulates and Axioms of the American School. In Allais, M. and Hagen,

[30] See the discussion found in White (2009), Elga (2010), Joyce (2010).

[31] For some examples, see Levi (1974), Gärdenfors and Sahlin (1982), and Joyce (2010).

O. (eds.) *Expected Utility Hypothesis and the Allais Paradox*. pp. 27–145. Dordrecht: Reidel. (Originally published in French.)
Armendt, B. (1986) A Foundation for Causal Decision Theory. *Topoi*. 5. 1. pp. 3–19.
Bartha, P. (2007) Taking Stock of Infinite Value: Pascal's Wager and Relative Utilities. *Synthese*. 154. 1. pp. 5–52.
Bermúdez, J. (2009) *Decision Theory and Rationality*. Oxford: Oxford University Press.
Bernoulli, D. (1738/1954) Exposition of a New Theory on the Measurement of Risk. *Econometrica*. 22. 1. pp. 23–36.
Bolker, E. (1965) *Functions Resembling Quotients of Measures*. Ph.D. dissertation for Harvard University.
Bolker, E. (1966) Functions Resembling Quotients of Measures. *Transactions of the American Mathematical Society*. 124. pp. 292–312.
Bolker, E. (1967) A Simultaneous Axiomatization of Utility and Subjective Probability. *Philosophy of Science*. 34. pp. 333–40.
Broome, J. (1991) *Weighing Goods: Equality, Uncertainty, and Time*. Oxford: Blackwell Publishers Ltd.
Broome, J. (1993) Can a Humean be moderate? In Frey, G. and Morris, C. (eds.) *Value, Welfare and Morality*. pp. 279–86. Cambridge: Cambridge University Press.
Buchak, L. (2013) *Risk and Rationality*. Oxford: Oxford University Press.
Christensen, D. (1991) Clever Bookies and Coherent Belief. *Philosophical Review*. 100. 2. pp. 229–47.
Christensen, D. (2001) Preference-Based Arguments for Probabilism. *Philosophy of Science*. 68. 3. pp. 356–76.
Colyvan, M. (2008) Relative Expectation Theory. *Journal of Philosophy*. 105. 1. pp. 37–44.
Davidson, D. (1973) Radical Interpretation. *Dialectica*. 27. 3-4. pp. 313–28.
de Finetti, B. (1937/1980) Foresight: Its Logical Laws, Its Subjective Sources. In Kyburg, H. and Smokler, H. (eds.) *Studies in Subjective Probability*. New York, NY: Robert E. Krieger Publishing Co.
Dreier, J. (1996) Rational Preference: Decision Theory as a Theory of Practical Rationality. *Theory and Decision*. 40. pp. 249–76.
Easwaran, K. (2008) Strong and Weak Expectations. *Mind*. 117. pp. 633–41.
Elga, A. (2010) Subjective Probabilities Should be Sharp. *Philosophers' Imprint*. 10. 5.
Ellsberg, D. (1961) Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*. 75. 4. pp. 643–69.
Fermat, P. and Pascal, B. (1654/1929) Fermat and Pascal on Probability. Translated from the French by Professor Vera Sanford. In Smith, D. (ed.) *A Source Book in Mathematics*. pp. 546–65. New York, NY: McGraw-Hill Book Co.
Fishburn, P. (1981) Subjective Expected Utility: A Review of Normative Theories. *Theory and Decision*. 13. pp. 139–99.
Gärdenfors, P. and Sahlin, N. (1982) Unreliable Probabilities, Risk Taking, and Decision Making. *Synthese*. 53. pp. 361–86.
Gibbard, A. and Harper, W. (1978/1981) Counterfactuals and Two Kinds of Expected Utility. In Harper, W., Stalnaker, R., and Pearce, G. (eds.) *Ifs: Conditionals, Belief, Decision, Chance, and Time*. pp. 153–90. Dordrecht: Reidel.
Gilboa, I. (1987) Expected Utility with Purely Subjective Non-Additive Probabilities. *Journal of Mathematical Economics*. 16. pp. 65–88.

Hammond, P. (1988) Consequentialist Foundations for Expected Utility. *Theory and Decision*. 25. pp. 25–78.
Hansson, B. (1988) Risk Aversion as a Problem of Conjoint Measurement. In Gärdenfors, P. and Sahlin, N. (eds.) *Decision, Probability, and Utility*. pp. 136–58. Cambridge: Cambridge University Press.
Hume, D. (1731/1978) *A Treatise of Human Nature*. (ed. L. A. Swilby-Bigge and P. Nidditch.) Oxford: Oxford University Press.
Hurley, S. (1989) *Natural Reasons, Personality and Polity*. New York, NY: Oxford University Press.
Jeffrey, R. (1965) *The Logic of Decision*. New York, NY: McGraw-Hill.
Jeffrey, R. (1983) *The Logic of Decision*. 2nd edition. Chicago, IL: University of Chicago Press.
Joyce, J. (1999) *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.
Joyce, J. (2010) A Defense of Imprecise Credences in Inference and Decision Making. *Philosophical Perspectives*. 24. 1. pp. 281–323.
Kahneman, D. and Tversky, A. (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*. 47. pp. 263–91.
Kolodny, N. (2007) How Does Coherence Matter? *Proceedings of the Aristotelian Society*. 107. pp. 229–63.
Kolodny, N. (2008) The Myth of Practical Consistency. *European Journal of Philosophy*. 16. 3. pp. 366–402.
Levi, I. (1974) On Indeterminate Probabilities. *Journal of Philosophy*. 71. pp. 391–418.
Levi, I. (1991) Consequentialism and Sequential Choice. In Bacharach, M. and Hurley, S. (eds.) *Foundations of Decision Theory*. pp. 92–122. Oxford: Basil Blackwell Ltd.
Lewis, D. (1974) Radical Interpretation. *Synthese*. 23. pp. 331–44.
Lewis, D. (1981) Causal Decision Theory. *Australasian Journal of Philosophy*. 59. pp. 5–30.
Machina, M. and Schmeidler, D. (1992) A More Robust Definition of Subjective Probability. *Econometrica*. 60. 4. pp. 745–80.
Maher, P. (1993) *Betting on Theories*. Cambridge: Cambridge University Press.
McClennen, E. (1990) *Rationality and Dynamic Choice: Foundational Explorations*. Cambridge: Cambridge University Press.
Meacham, Christopher J. G. and Weisberg, Jonathan (2011) Representation Theorems and the Foundations of Decision Theory. *Australasian Journal of Philosophy*. 89. 4. pp. 641–63.
Nover, H. and Hájek, A. (2004) Vexing Expectations. *Mind*. 113. pp. 237–49.
Nozick, Robert (1969) Newcomb's Problem and Two Principles of Choice. In Rescher, Nicholas (ed.) *Essays in Honor of Carl G. Hempel*. pp. 114–15. Synthese Library. Dordrecht: Reidel
Okasha, S. (2007) Rational Choice, Risk Aversion and Evolution. *Journal of Philosophy*. 104. 5. pp. 217–35.
Pettit, P. (1991) Decision Theory and Folk Psychology. In Bacharach, M. and Hurley, S. (eds.) *Foundations of Decision Theory*. pp. 147–75. Oxford: Basil Blackwell Ltd.
Ramsey, F. (1926/1931) Truth and Probability. In Braithwaite, R. (ed.) *The Foundations of Mathematics and other Logical Essays*. London: Kegan, Paul, Trench, Trubner & Co. (Published posthumously)
Samuelson, P. (1938) A Note on the Pure Theory of Consumer's Behaviour. *Economica*. 5. 17. pp. 61–71.

Savage, L. (1954/1974) *The Foundations of Statistics*. 2nd edition. Dover: John Wiley and Sons, Inc.

Schmidt, U. (2004) Alternatives to Expected Utility: Formal Theories. In Barberà, S., Hammond, P., and Seidl, C. (eds.) *Handbook of Utility Theory*. pp. 757–837. Boston, MA: Kluwer Academic Publishers.

Seidenfeld, T. (1988) Decision Theory Without 'Independence' or Without 'Ordering'. *Economics and Philosophy*. 4. pp. 267–90.

Skyrms, B. (1982) Causal Decision Theory. *The Journal of Philosophy*. 79. 11. pp. 695–711.

Spohn, W. (1977) Where Luce and Krantz do Really Generalize Savage's Decision Model. *Erkenntnis*. 11. pp. 113–34.

Stalnaker, R. (1972/1981) Letter to David Lewis. In Harper, W., Stalnaker, R., and Pearce, G. (eds.) *Ifs: Conditionals, Belief, Decision, Chance, and Time*. pp. 151–2. Dordrecht: Reidel.

Sugden, R. (2004) Alternatives to Expected Utility: Foundations. In Barberà, S., Hammond, P., and Seidl, C. (eds.) *Handbook of Utility Theory*. pp. 685–755. Boston, MA: Kluwer Academic Publishers.

Vallentyne, P. (1993) Utilitarianism and Infinite Utility. *Australasian Journal of Philosophy*. 71. 2. pp. 212–17.

von Neumann, J. and Morgenstern, O. (1944) *Theory of Games and Economic Behavior*. Princeton, NJ: Princeton University Press.

Wakker, P. (1988) Nonexpected Utility as Aversion to Information. *Journal of Behavioral Decision Making*. 1. pp. 169–75.

White, R. (2009) Evidential Symmetry and Mushy Credence. In Gendler, T. and Hawthorne, J. (eds.) *Oxford Studies in Epistemology*. pp. 161–86. Oxford: Oxford University Press.

Zynda, L. (2000) Representation Theorems and Realism about Degrees of Belief. *Philosophy of Science*. 67. 1. pp. 45–69.

CHAPTER 37

PROBABILISTIC CAUSATION

CHRISTOPHER HITCHCOCK

37.1 Introduction

This chapter will explore a variety of projects that aim to characterize causal concepts using probability. I will somewhat arbitrarily divide these into four categories. First, I will discuss a tradition within philosophy that has aimed to define, or at least constrain, causation in terms of conditional probability. Secondly, I will discuss the use of causal Bayes nets to represent causal relations, to facilitate inferences from probabilities to causal relations, and to 'identify' causal quantities in probabilistic terms. Thirdly, I will discuss efforts to measure causal strength in probabilistic terms, with particular attention to the significance of these measures in the context of epidemiology. Finally, I will discuss attempts to analyze the relation of 'actual causation' (sometimes called 'singular causation') using probability.

37.2 Probabilistic Theories of Causation

Hans Reichenbach (1956) was the first philosopher to attempt to define causation in probabilistic terms. Other important contributions to this program have included Suppes (1970), Cartwright (1979), Skyrms (1980), and Eells (1991). I will not attempt to trace this history here, nor to discuss the differences between these authors. (See Hitchcock (2010) for a detailed discussion.) Instead I will present a composite probabilistic theory of causation that can be used to illustrate some of the key ideas.

The central idea behind probabilistic theories of causation is that causes need not be necessary or sufficient for their effects. For example, the claim that prolonged exposure to formaldehyde causes nasal cancer does not imply that all or only those exposed to formaldehyde develop nasal cancer. Instead, a cause need only raise the probability of its effect. The natural way to capture this idea is in terms of conditional probability:

$$P(E|C) > P(E|{\sim}C).$$

C and E are events (in the probability-theoretic sense) that represent the cause and the effect, respectively. There are several problems with defining causation directly in this way. First, this relation of probability-raising is symmetric in C and E. If C raises the conditional probability of E, then E will raise the conditional probability of C. Causation, by contrast, is usually taken to be asymmetric: if C causes E, then E does not cause C. One way to deal with this problem would be to add a time index to the events in the probability space, and to require that causes precede their effects in time. (While Lewis (1979/1986) and others have worried that this approach rules out backwards-in-time or simultaneous causation by fiat, we will not address this concern here.)

A second problem is that two events can be correlated owing to the presence of a common cause. Jeffrey (1969) gives the following example: When the reading on a barometer drops below a certain level (event B), a storm (S) frequently follows. In particular, the following inequality holds: $P(S|B) > P(S|{\sim}B)$. However, the drop in the barometer reading does not cause the storm; rather, both of these events are caused by a drop in atmospheric pressure (A). Reichenbach (1956) claimed that in cases such as this, the common cause *screens off* its two effects. Formally, A screens B off from S just in case $P(S|BA) = P(S|{\sim}BA)$. Reichenbach also claimed that the absence of the cause (${\sim}A$) screens off the two effects: $P(S|B{\sim}A) = P(S|{\sim}B{\sim}A)$. Given a drop in atmospheric pressure, a storm is no more likely to occur when the barometer reading drops than when it doesn't (perhaps due to a malfunction); similarly, given no drop in pressure, the barometer reading makes no difference to the chance of a storm.

In Jeffrey's example, the presence of a common cause produces a probabilistic correlation between B and S, even though neither event causes the other. In such a case, we say that the correlation is *spurious*. The common cause A is sometimes called a *confounder* or a *latent cause*. It is also possible for a common cause to have the opposite effect, masking a causal relation between two events. For instance, vaccination against yellow fever prevents yellow fever (i.e., causes one to remain fever-free). However, people typically get vaccinated for yellow fever only if they live in or travel to a region where yellow fever is endemic. Thus, it may be that people who are vaccinated against yellow fever suffer from yellow fever in the same proportion overall as those who are not vaccinated. In this case, the correlation between vaccination and being fever-free should appear once we condition on the common cause: living in or traveling to a yellow fever zone.

Reichenbach (1956) also pointed out a different kind of case where screening-off relations occur. If a distal cause A causes an effect E via an intermediate cause C, and through no separate causal process, then the intermediate cause C will screen off the distal cause from the distal effect: $P(E|AC) = P(E|{\sim}AC)$ and $P(E|A{\sim}C) = P(E|{\sim}A{\sim}C)$. For example, if intravenous drug use causes AIDS by causing HIV infection, then HIV infection will screen off intravenous drug use from AIDS.

Laying out these ideas more precisely, suppose that we have a population p of individuals. It could be a population of humans (such as the population of Great Britain), or of animals or objects, or it could be a sequence of experiments or similar events. Let $\mathbf{K}$ be a set of time-indexed properties of the form A_t. $A_t(i)$ asserts that individual i has property A at time t. We will assume that the properties are all logically and mereologically distinct (they do not stand in logical or part-whole relations to one another). The time indices do not indicate absolute times (e.g. dates and years), but rather times that are relative to the individuals in the population. For instance, $\mathbf{K}$ might include events F_{40} and N_{50}, representing the

properties of being exposed to formaldehyde at age 40 and developing nasal cancer at age 50 (as opposed to being exposed to formaldehyde in 1940 and developing nasal cancer in 1950). The population p will be assumed to be sufficiently well defined to determine a range of possible properties, and probabilities that an individual will have these properties at various times. We assume that the probability function P is defined on the Boolean algebra generated by the time-indexed properties in **K**.

A primitive relation of *causal relevance* holds among the time-indexed properties. For instance, in a particular population, exposure to formaldehyde at age 40 might be causally relevant to developing nasal cancer at age 50. C_t will be causally relevant to $E_{t'}$ only if t is earlier than t'. C_t may be causally relevant to $E_{t'}$ by causing it, preventing it, or affecting it in some other way. The negations of these properties, representing the absences of the properties in individuals at the relevant times, can also stand in the relation of causal relevance. In particular, if C_t is causally relevant to $E_{t'}$, $\sim C_t$ will be causally relevant to $E_{t'}$ and C_t will be causally relevant to $\sim E_{t'}$. (If C_t causes $E_{t'}$, $\sim C_t$ will prevent $E_{t'}$ and C_t will prevent $\sim E_{t'}$, and analogously for other types of causal relevance.)

In order to evaluate the causal relevance of C_t for $E_{t'}$ we must construct a partition $\{B_1, \ldots, B_n\}$ of *background* conditions. We do this by taking the subset $\mathbf{K}'$ of **K** such that:

$A_{t''} \in \mathbf{K}'$ iff

(I) $A_{t''}$ is not identical to C_t or $\sim C_t$;
(II) $A_{t''}$ is causally relevant to $E_{t'}$; and
(III) C_t is not causally relevant to $A_{t''}$.

If $A_{t''}$ satisfies these conditions, $\sim A_{t''}$ will as well. Thus $\mathbf{K}'$ will be a union of pairs of the form $\{A_{t''}, \sim A_{t''}\}$. Then each background condition B_i will be a maximally specific conjunction of consistent members of $\mathbf{K}'$. That is, for each pair $\{A_{t''}, \sim A_{t''}\}$, one or the other will appear as a conjunct in each B_i. Intuitively, each B_i holds fixed the presence or absence of each property in $\mathbf{K}'$.

We then define:

(i) C_t *is a cause of* $E_{t'}$ iff $t < t'$ and for all B_i, $P(E_{t'}|C_tB_i) > P(E_{t'}|\sim C_tB_i)$;
(ii) C_t *prevents* $E_{t'}$ iff $t < t'$ and for all B_i, $P(E_{t'}|C_tB_i) < P(E_{t'}|\sim C_tB_i)$;
(iii) C_t is an *interacting cause* of $E_{t'}$ iff $t < t'$, C_t does not cause or prevent $E_{t'}$, and for some B_i, $P(E_{t'}|C_tB_i) \neq P(E_{t'}|\sim C_tB_i)$.

If C_t is an interacting cause of $E_{t'}$ in a population p, it may be a cause or preventer of $E_{t'}$ in some subpopulation of p in which individuals with certain properties in $\mathbf{K}'$ are excluded.

While the primitive notion of causal relevance was undefined, we impose the following constraint:

(iv) C_t is causally relevant to $E_{t'}$ iff C_t is a cause, preventer, or interacting cause of $E_{t'}$.

In general, the probability distribution over (Boolean combinations of) the properties in K will not uniquely determine the causal relations among these properties, but it will rule out some systems of causal relations.

The normal experimental practice of randomization can be understood as distributing each of the background conditions B_i evenly between the treatment (C_t) condition and the control ($\sim C_t$) condition. Thus a randomized experiment will yield the conclusion that C_t is a cause of $E_{t'}$ when

(R) $\Sigma_i P(E_{t'}|C_t B_i)P(B_i) > \Sigma_i P(E_{t'}|\sim C_t B_i)P(B_i)$.

In practice, of course, the statistics generated by an experiment may not conform exactly to the underlying probabilities, and it may be difficult to determine whether (R) holds when the sample size and probability differences are relatively small.

This program in probabilistic causation made important advances in our understanding of the connection between causal relationships and probabilistic screening-off relations. These connections form the core of the newer approaches to causal modeling that will be the focus of the next section. Nonetheless, the program has a number of shortcomings.

First, as presented here, the probabilistic theory of causation does two different things: (a) it imposes probabilistic constraints upon the relation of causal relevance that is treated as a primitive within the theory; (b) it uses probability together with the primitive notion of causal relevance to classify different kinds of causal relevance – causation, prevention, and interaction. That is not in itself a problem. But while the theory was being developed, most writers did not clearly distinguish between these two projects, and even conflated them. In particular, it was widely assumed that causation, in the sense of probability-raising, was the primary target of philosophical analysis. This led to a number of debates about how best to define causation: whether to require that causes raise the probability of their effects in *all* background conditions, as proposed in (i) above (and defended by Cartwright (1979) and Eells (1991)); whether causes must raise the probability of their effects in some background conditions and lower it in none (a Pareto-dominance condition proposed by Skyrms (1980)); or whether causes must raise the probability of their effects on average, as suggested by (R) above (a position defended by Dupré (1984)). In hindsight, this appears to be just a debate about how best to divide the relation of causal relevance into different categories, rather than a substantial debate about the metaphysics of causation.

A related point is that there are different reasons for holding factors fixed, and these get run together in the type of theory outlined here. For example, if we want to know whether C_t has an interactive effect on $E_{t'}$ we will need to hold fixed all independent causes of $E_{t'}$, even those that are not also causes of C_t. Consider a variant on an example of Cartwright (1979). Ingestion of an acid poison (A_t) increases the probability of death ($D_{t'}$). However, in the rare circumstance in which a base poison (B_t) is also present, ingestion of an acid poison lowers the probability of death. Thus ingestion of an acid poison is an interacting cause of death. However, if the question is whether the correlation between ingestion of an acid poison and death is spurious, it is only necessary to hold fixed those factors that are common causes of both. In this case, it is not necessary to hold fixed the presence or absence of a base poison.

Secondly, the theory requires that we control for *all* causes of $E_{t'}$ that are not caused by C_t. In practice, we never have access to a complete list of causes of $E_{t'}$. Nothing in the theory tells us when we can make do with some subset of the causes of $E_{t'}$, when we have 'enough' causes to be able to apply the theory and get correct results. As a result, despite its superficial

connection to statistical and experimental methods for inferring causation, the theory does not provide much by way of usable rules of causal inference.

37.3 Causal Models

I will describe here just one kind of causal model, the causal Bayes net. A causal Bayes net uses a directed acyclic graph (DAG) to represent causal relations among a set of variables, and pairs it with a probability distribution over the set of variables. The origin of this approach can be traced back to work by Sewall Wright (1921), but it has been developed extensively by Spirtes, Glymour, and Scheines (2000) and Pearl (2009).

A causal Bayes net is a triple $\langle \mathbf{V}, \mathbf{G}, P \rangle$, where $\mathbf{V}$ is a set of variables, $\mathbf{G}$ is a directed acyclic graph over $\mathbf{V}$, and P is a probability distribution over the field of events generated by $\mathbf{V}$. The variables in $\mathbf{V}$ correspond to the properties of an individual in some population. For example, in a population of American adults, we might have variables representing an individual's education level, work experience, and present income. A variable can be binary, representing the presence or absence of some property, but it can also be multiple-valued or continuous.

$\mathbf{G}$ is a set of ordered pairs of variables in $\mathbf{V}$. If $\langle X, Y \rangle \in \mathbf{G}$, we represent this graphically by drawing an arrow from X to Y, and we say that X is a *parent* of Y. If there are arrows from X_1 to X_2, X_2 to X_3,..., and X_{n-1} to X_n, then there is a *directed path* from X_1 to X_n. In this case, we say that X_1 is an *ancestor* of X_n, and that X_n is a *descendant* of X_1. $\mathbf{G}$ is *acyclic* if there is no directed path from any variable to itself. $\mathbf{PA}(X)$ is the set of all parents of X and $\mathbf{ND}(X)$ is the set of all variables in $\mathbf{V}$ other than X and its descendants. Figure 37.1 shows a DAG on the variable set $\{R, S, T, U, W, Z\}$. A DAG represents the qualitative causal structure among the set of variables in an individual from the relevant population. In particular, an arrow from X to Y indicates that X has a causal influence on Y that is not mediated by any other variables in the set. In this case, we say that X is a *direct cause* of Y. In figure 37.1, R is a direct cause of S, but only an indirect cause of Z. An arrow says nothing about the nature of the causal influence. For example, it may be that larger values of R tend to produce larger values of S (so R and S are positively correlated), or that larger values of R produce smaller values of S (they are negatively correlated), or that some more complicated relationship holds.

The probability distribution P defined on the field of events generated by the variables in $\mathbf{V}$ satisfies the *Markov Condition* with respect to the DAG $\mathbf{G}$ just in case:

(MC) For every $X \in \mathbf{V}$, and every $\mathbf{Y} \subseteq \mathbf{ND}(X)$:

$$P(X \,|\mathbf{PA}(X), \mathbf{Y}) = P(X \,|\mathbf{PA}(X)).^{1}$$

That is, the values of the parents of X screen the values of X off from the values of all variables except for X and its descendants. For example, suppose that P satisfies the Markov Condition

[1] This notation, in which variables and sets of variables appear within the conditional probabilities, is shorthand for: For every H, $\boldsymbol{H_1}$, and $\boldsymbol{H_2}$ such that $P(X \in H \,|\mathbf{PA}(X) \in \boldsymbol{H_1}, \mathbf{Y} \in \boldsymbol{H_2})$ is defined, $P(X \in H \,|\mathbf{PA}(X) \in \boldsymbol{H_1}, \mathbf{Y} \in \boldsymbol{H_2}) = P(X \in H | \mathbf{PA}(X) \in \boldsymbol{H_1})$.

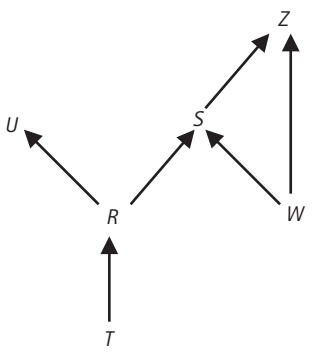


FIGURE 37.1 A directed acyclic graph.

with respect to the graph shown in figure 1. Then R screens off U from S, R screens off S from T, and R and W will be probabilistically independent (they are screened off by the null set of variables); there will be additional cases of screening-off as well. Some of the screening-off relations implied by MC are of the kinds described earlier by Reichenbach (1956). If P satisfies the MC with respect to $\mathbf{G}$, then $\langle \mathbf{V}, \mathbf{G}, \mathrm{P} \rangle$ is a *Bayes net*. Richard Neapolitan and Xia Jiang's chapter (10) in this volume includes descriptions of many of the formal properties of Bayes nets, as well as computation using Bayes nets.

Suppose the variables in $\mathbf{V}$ represent a range of variables occurring in a population, $\mathbf{G}$ represents the actual causal relations among these variables, and P describes the empirical probability distribution over the values of the variables. We do not automatically assume that P satisfies the Markov Condition with respect to $\mathbf{G}$. For example, suppose that $\mathbf{V} = \{R, S\}$, neither variable causes the other, but they share a common cause Z that is not in $\mathbf{V}$. Then R and S will be correlated, when the Markov Condition says they should be independent. More generally, we will say that a variable set $\mathbf{V}$ is *causally insufficient* if there exist X, $Y \in \mathbf{V}$, and a variable $Z \notin \mathbf{V}$, such that if Z were added to $\mathbf{V}$, it would be a direct cause of both X and Y. Causally insufficient sets of variables are not expected to satisfy the Markov Condition.

The *Causal Markov Condition* (CMC) is the assumption that for any suitable set of variables $\mathbf{V}$, the empirical probability distribution P over $\mathbf{V}$ will satisfy the MC with respect to $\mathbf{G}$, the DAG representing the causal relations among the variables in $\mathbf{V}$. 'Suitability' here implies causal sufficiency, and perhaps other conditions as well. In this case, we say that $\langle \mathbf{V}, \mathbf{G}, \mathrm{P} \rangle$ is a *causal Bayes net*. Even if we don't assume that $\langle \mathbf{V}, \mathbf{G}, \mathrm{P} \rangle$ is itself a causal Bayes net, we can often infer a great deal from the assumption that the probability distribution P is generated by some richer structure that does have the properties of a causal Bayes net.

One justification for assuming the Causal Markov Condition is a theorem due to Pearl and Verma (1991). Suppose that we have a variable set $\mathbf{V}$, and DAG $\mathbf{G}$ representing the causal relations among the variables in $\mathbf{V}$. Suppose, in addition, that the value of each variable X in $\mathbf{V}$ is a deterministic function of its parents in $\mathbf{V}$, together with an 'error variable' E_X, which represents the influence of any variables that are not included in $\mathbf{V}$. In other words, $X = f_X(\mathbf{PA}(X), E_X)$. Then the values of all of the error variables will uniquely determine the value of all of the variables in $\mathbf{V}$. Thus a probability distribution P' over the error variables will induce a probability distribution P over the variables in $\mathbf{V}$. If the error variables are independent in P', then the induced probability distribution P will satisfy MC with respect

to **G**. The idea is that if we include enough variables in **V** so that any remaining causal influences are probabilistically independent, then the probability distribution over **V** will satisfy MC.

While the Causal Markov Condition entails that certain conditional independence relations will hold, it does not imply that any (conditional or unconditional) dependence relations hold. For instance, a probability distribution that made {*R, S, T, U, W, Z*} all independent (in any combination) would satisfy the MC with respect to the graph in figure 37.1. Thus it is necessary to supplement the CMC with other conditions. Spirtes, Glymour, and Scheines (2000) propose the *Minimality Condition*:

(Min) No subgraph of **G** satisfies the Markov Condition with respect to *P*.

For example, in figure 37.1, if *P* satisfies Min, then *P* will not satisfy MC with respect to the graph that results from removing the arrow from *W* to *Z*. This means that *S* does not screen off *W* from *Z*. This helps to give meaning to the arrow from *W* to *Z*: it implies that *Z* probabilistically depends upon *W* when we condition on (some value of) *S*.

Pearl (1988, chapter 3) proves a theorem that, suitably interpreted, shows that causation can be reduced to probability and time order under certain conditions. Suppose that the variables in **V** are time-indexed, so that earlier variables can cause later ones but not vice versa. In addition, suppose that every possible combination of values of the variables in **V** has positive probability. Then there is only one graph **G** that is compatible with MC and Min. If time order is not given, however, it may not be possible to recover the structure of the graph **G** from probabilities alone.

In causal inference problems, it is common to assume a condition that is strictly stronger than Minimality. Spirtes, Glymour, and Scheines (2000) call this the *Faithfulness Condition*:

(F) There are no (conditional or unconditional) independence relations that are not implied by the Markov Condition.

For example, in figure 37.1, there are two directed paths from *W* to *Z*. This indicates that *W* influences *Z* in two different ways: it has a direct influence on *Z*, and it also influences *Z* via *S*. It is, in principle, possible that these two influences exactly cancel each other out. In that case, *W* and *Z* would be (unconditionally) independent. This would not violate (Min), but it would violate (F). The Faithfulness Condition is usually thought of as a methodological assumption, rather than a metaphysical one. It is not impossible, or contrary to the definition of causation, for the two causal influences from *W* to *Z* to exactly cancel each other out. Rather, the assumption says that when we encounter a probabilistic independence (either conditional or unconditional), we should prefer a causal structure that *entails* the independence to one which is merely *compatible* with the independence.

Let $\langle \mathbf{V}, \mathbf{G}, P \rangle$ be a causal model, with $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$, and where *P* satisfies the Markov Condition with respect to **G**. Then the joint probability distribution factorizes in a simple way:

$$P(X_1, X_2, \ldots, X_n) = \Pi_i P(X_i | \mathbf{PA}(X_i)).$$

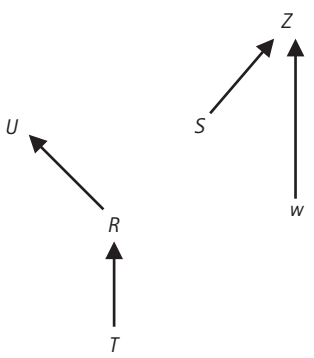


FIGURE 37.2 A new directed acyclic graph, representing the result of an intervention on the variable *S* in figure 37.1.

This means that all of the probabilities assigned by P can be derived from the causal probabilities, that is, the probabilities conferred on each variable X_i by its direct causes.

If we *intervene* on a variable X_i, we impose a new probability distribution $\mathrm{P}'(X_i)$ on the values of the variable, and make it independent of its normal causes. This is what we do (or hope to do) when we conduct a randomized experiment. For example, if we have a pool of experimental subjects, and toss a coin to determine whether each subject receives the trial drug or a placebo, we impose a probability of .5 for each of these possibilities, and make them independent of their usual causes. It may be that, in the population at large, whether or not an individual has health insurance influences whether or not she takes the drug. But in the randomized trial, health insurance no longer influences whether or not someone takes the drug.[2] The result of performing such an intervention will be a new probability distribution:

$$P'(X_1, X_2, \ldots, X_n) = P'(X_i)\ \Pi_{j \neq i} P(X_j | \mathbf{PA}(X_j)).$$

The new probability distribution P' will satisfy the MC with respect to the graph $\mathbf{G}'$ that results from removing the arrows into X_i. The intervention removes the causal influences on X_i and imposes a probability distribution on X_i directly. From the definition of MC it follows that in the new distribution X_i will be probabilistically independent of all other variables except for its descendants. This captures the conventional wisdom that randomized trials provide the best framework for inferring causal relationships. Figure 37.2 shows the result of an intervention on the variable *S* in figure 37.1.

One special case of such an intervention occurs when we set the value of X_i to a specific value. This is the case where P' assigns probability one for that value of X_i, and probability zero to all others. A non-backtracking counterfactual, in the sense of Lewis (1979/1986), can be represented in this way. Interventions upon multiple variables can also be represented in the same manner.

Causal models can be used to address many different kinds of problem. These can be divided into two broad categories. The first category is *causal inference* or *causal discovery*. Here, the goal is to infer the qualitative causal structure among the set of variables in **V**, i.e., to infer the graph **G** on **V**. The structure of the problem will vary depending upon two things. First, what assumptions do we make about the true causal model $\langle \mathbf{V}, \mathbf{G}, P \rangle$? Do we

[2] Woodward (2003) has an extended discussion of the properties such an intervention must have.

assume that **V** is causally complete, so that P will satisfy the MC with respect to **G**? Or do we allow for the possibility that there may be latent causes? Do we assume the Faithfulness Condition, or perhaps some weaker condition?[3] Secondly, what kind of information do we have available? Do we have an observed probability distribution on all of the variables in **V**? Have we observed the results of one or more interventions on variables in **V**?[4] Do we have constraints upon the possible causal structures, perhaps from time order information, or from background theory? In general, it may not be possible to infer the unique graph **G** that describes the causal relations among the variables in **V**, but only to infer that the correct graph is among some class of possible graphs that are all consistent with the assumptions, constraints, and information that we have.

The second class of problems comprises *identification* problems, where the goal is to compute the value of a causally significant quantity. For instance, the goal might be to compute the probability distribution on the variable X_j that would result from an intervention that sets the value of X_i to some specific value. These problems are especially important in fields that are policy-oriented, such as epidemiology. For instance, we might be interested in predicting how the incidence of influenza would change if an immunization program was implemented. Again, the structure of the problem will vary depending upon the assumptions that are made, and the information that is available. Do we know the correct qualitative causal structure, or have partial information about it? Have we observed the results of other interventions? Have we observed the probability distribution over all of the variables? Etc.

37.4 Measures of Causal Strength

Probabilistic approaches to causation lend themselves naturally to measures of *causal strength*. Consider the case where we have a binary cause C and effect E. Let $P_C(E)$ be the probability for E that would result from an intervention that brings about C, and $P_{\sim C}(E)$ be the probability that would result from an intervention that brings about the absence of C. Since we are assuming that we are intervening on C (e.g. by randomly determining whether or not C is present), any difference between $P_C(E)$ and $P_{\sim C}(E)$ will be due to the causal influence of C on E, and not due to spurious correlation. It is natural to judge that C is a strong cause of E to the extent that $P_C(E)$ is much larger than $P_{\sim C}(E)$. But how are we to compare these probability values? Do we look at the difference $P_C(E) - P_{\sim C}(E)$, the ratio $P_C(E)/P_{\sim C}(E)$, or some other function of the probabilities? It should be obvious that the ratio can be large while the difference is relatively small, and vice versa. These issues have been discussed at some length in the field of epidemiology, which is often concerned with how great a risk a certain infectious agent or environmental hazard is.

[3] There will typically be further assumptions as well. For instance, it is usually assumed that there is no *inter-unit* causation. This can fail, for example, if the population includes family members who live together, and one person's smoking causes another's lung cancer. There will also be general inductive or statistical assumptions involved in inferring the probability distribution P from an observed sample.

[4] For an introduction to the topic of causal discovery using interventions, see Eberhardt and Scheines (2007).

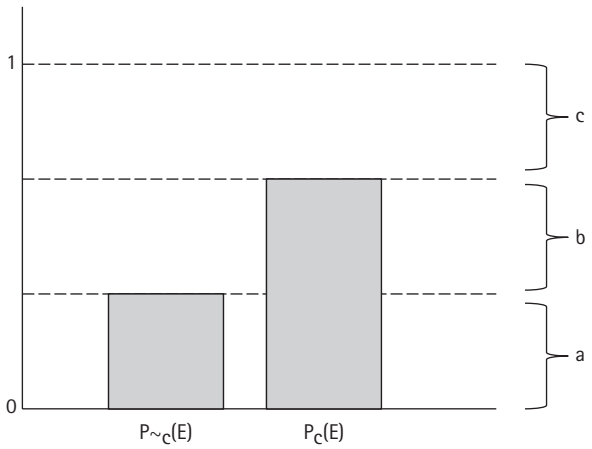


FIGURE 37.3 The [0, 1] interval is divided into three subintervals, *a*, *b*, and *c*, corresponding to $P_{\sim C}(E)$, $P_C(E) - P_{\sim C}(E)$, and $1 - P_C(E)$, respectively.

These issues can also arise in a legal context, where it can be important to judge how likely it is that somebody's illness was caused by exposure to a particular hazard. Terminology varies widely; for definiteness, I will use the terminology recommended by Rothman et al. (2008).

It may be helpful to picture these measures of causal strength visually. Figure 37.3 shows a simple bar graph, comparing $P_C(E)$ and $P_{\sim C}(E)$, where $P_C(E) > P_{\sim C}(E)$. The interval [0, 1] on the vertical axis is divided into three intervals, *a*, *b*, and *c*. I will use the same letters to denote the size of these intervals. Thus $P_{\sim C}(E) = a$, $P_C(E) = a + b$, and $1 = a + b + c$.

The typical situation in epidemiology is that *C* is an infectious agent, toxin, or other potential cause of disease or harm, denoted by *E*. The probability of *E* is referred to as *risk*, a term which carries a negative connotation. The *causal risk difference* is the difference between the risk of *E* for those who are exposed to *C*, and for those who are not:

(RD) $P_C(E) - P_{\sim C}(E)$

This difference is equal to *b* in figure 3. The risk difference will lie in the [-1, 1] interval, with a negative number when *C* prevents *E*, a positive number when *C* causes *E*, and zero when *C* is neutral or irrelevant to *E*. The *causal risk ratio* or the *relative risk* is the ratio between the same two quantities:

(RR) $P_C(E)/P_{\sim C}(E)$

This is equal to $(a + b)/a$ or $1 + b/a$ in figure 37.3. The risk ratio will lie in the interval [0, ∞], with values less than one for prevention, greater than one for causation, and equal to one for neutrality.

The risk difference divided by $P_C(E)$ is called the *excess fraction*:

(EF) $[P_C(E) - P_{\sim C}(E)]/P_C(E)$

This is equal to $b/(a + b)$ in figure 3. A little algebra reveals that the excess fraction is also equal to 1 – 1/RR, where RR is the risk ratio $P_C(E)/P_{\sim C}(E)$. It follows that the excess fraction is *ordinally equivalent* to the risk ratio: the excess fraction of C for E is greater than or equal to the excess fraction of C' for E' just in case the risk ratio of C for E is greater than or equal to the risk ratio of C' for E'. That is, the excess fraction and the risk ratio will agree in all their comparative judgments of causal strength. When C causes E, the excess fraction will have a value in (0, 1], just like the risk difference. In the case where C prevents E, it is more normal to talk about the excess fraction of C for $\sim E$, which will be in (0, 1], rather than the excess fraction of C for E, which would be in the interval $[-\infty, 0)$. For example, we might talk about the efficacy of a vaccine (C), in terms of its excess fraction for $\sim E$ (the absence of the disease).

The excess fraction is sometimes called the *attributable fraction*. The interpretation is roughly as follows: Among those in the population who were exposed to C, and suffered adverse effect E, what fraction of occurrences of E are *attributable* to C? This fraction is attributable to C in the sense that it is the proportion of cases that are in excess of those that would have occurred had C been absent.

The excess fraction is sometimes also called the *probability of causation*. The idea is that if an individual has been exposed to C, and suffered effect E, the excess fraction is the probability that E was caused by C. This quantity is of particular interest in tort law. In American tort law, the standard of evidence for a plaintiff to receive compensation is 'more probable than not'. Thus if a plaintiff has been exposed to hazard C, and suffered consequence E, he must convince a judge or jury that it is more probable than not that C caused E. This has often been interpreted as requiring a probability of causation greater than .5. If we equate probability of causation with excess fraction, this corresponds to a relative risk of 2. That is, in order for it to be more probable than not that C caused E, $P_C(E)$ must be at least two times greater than $P_{\sim C}(E)$: $P_C(E) \geq 2P_{\sim C}(E)$.[5]

However, the excess fraction can only be properly interpreted as the probability of causation under very strong assumptions. In particular, this interpretation assumes that the population can be divided into three types of individuals. *Doomed* individuals will suffer E regardless of whether they are exposed to C. These individuals make up a proportion $P_{\sim C}(E)$ of the population. (We can think of these individuals as lying in interval a in figure 37.3). *Susceptible* individuals will suffer E just in case they are exposed to C. They make up proportion $P_C(E) - P_{\sim C}(E)$ of the population (interval b in figure 37.3). Finally, *immune* individuals will avoid E come what may (proportion $1 - P_C(E)$, interval c in figure 37.3). Then, among those that are exposed to C and suffer E, C caused E in all and only those who were susceptible, but not doomed. (This would be the verdict of an analysis that identified causation with counterfactual dependence.[6]) These assumptions can fail for a variety of

[5] One important legal case where this standard was rejected was *Herskovits v. Group Health Cooperative of Puget Sound*, heard by the Supreme Court of Washington State in 1983 (664 P 2d 474). In that case, it was estimated that a physician's failure to diagnose lung cancer resulted in a patient's having a 25% chance of survival instead of a 39% chance. The patient died. The physician's failure did not double the risk of death (it increased it from 61% to 75%). The court ruled that a physician should not automatically be free of liability for negligence any time a patient has less than a 50% chance of survival.

[6] Lewis (1973/1986) famously propounded a counterfactual theory of causation, but he did not identify causation with counterfactual dependence. He took the latter to be sufficient, but not necessary for causation.

reasons. (i) C may prevent E in some individuals that would have suffered E in the absence of C. In this case, C must cause a corresponding additional number of cases of E in order to make up the full difference between $P_C(E)$ and $P_{\sim C}(E)$. (ii) C may accelerate the onset of E in individuals who would have eventually suffered E in the absence of C. (See e.g. Greenland and Robins (1988)) (iii) C may *preempt* the causes that would have brought about E in C's absence. (We will discuss preemption in greater detail in the next section.) (iv) C may causally contribute to the occurrence of E in those who would have suffered E anyway. ((ii) and (iii) are really just special cases of (iv).) (v) If the mechanisms producing E are *indeterministic*, it makes no sense to divide the population of individuals exposed to C into those who are doomed, susceptible, and immune. For any individual that was exposed to C and suffered E, there is no fact of the matter as to whether she would have suffered E in the absence of C. Perhaps all we can say is that she would have had some lower probability of suffering E. In this case, Lewis (1986) and Parascandola (1996) argue that we should count C as a cause of E in all of the cases where C increased the probability of E, and E occurred. Taking all of these possibilities into account, the most that we can say is that the excess fraction provides a *lower bound* on the probability of causation, properly speaking. That is, if e is the excess fraction, and p the probability of causation, the most we can infer is that $e \leq p \leq 1$. It is impossible to know the true probability of causation without more detailed information about the mechanisms responsible for producing E. Rothman et al. (2008) propose using *etiological fraction* for this true probability of causation, to distinguish it from the excess fraction.

Another important quantity is what Sheps (1958) called the *relative difference*:

(RelD) $[P_C(E) - P_{\sim C}(E)]/[1 - P_{\sim C}(E)]$.

This is equal to $b/(b + c)$ in figure 37.3. The psychologist Patricia Cheng has used the term *causal power* to describe this quantity,[7] and employs it in a theory of causal reasoning. The idea is that this quantity reflects the potential for C to cause E in those individuals who would not otherwise suffer E. Khoury et al. (1989) describe this in terms of the *susceptibility* of the population to C. When $P_{\sim C}(E)$ is high, the relative difference can still be large, even though the risk difference and risk ratio must be small. (The risk difference cannot exceed $1 - P_{\sim C}(E)$, and the risk ratio cannot exceed $1/P_{\sim C}(E)$.) However, these causal interpretations of the relative difference are subject to many of the same caveats discussed in the previous paragraph. In addition, this interpretation assumes that the conditions that make one susceptible to C are probabilistically independent of the conditions that bring about E in C's absence. (For the derivation of (RelD) using these assumptions, see Khoury et al. (1989) or Rothman et al. (2008, p. 65).)

A final quantity of interest is the *population-attributable fraction*:

(PA) $P(E) - P_{\sim C}(E)$.

Unlike the other measures, this one involves the unconditional probability of E. This will depend in turn upon the unconditional probabilities of C and $\sim C$. For instance, if C raises

[7] More precisely, the causal power reduces to this quantity under specific conditions. See Cheng (1997) for detailed discussion.

the probability of E, then the population-attributable fraction will be larger when $P(C)$ is greater. This quantity is often thought of as the amount of E in the population that is due to C. Thus, this quantity will be larger when C is more prevalent. More immediately, the population-attributable fraction can sometimes be interpreted as the amount by which the prevalence of E would decrease if an intervention were to eliminate all exposure to C in the population. This interpretation assumes that there is no *inter-unit* causation.[8]

Detailed discussion of the mathematical properties of these and other measures is contained in Fitelson and Hitchcock (2011).

37.5 Actual Causation

Many philosophers and legal theorists have been interested in the relation of *actual causation*. Relations of actual causation are described by sentences, typically expressed in the past tense, asserting that one particular event was caused by another. For example, if Billy and Suzy throw rocks at a window, and Suzy's rock shatters the window while Billy's misses, we would say that Suzy's throw caused the window to shatter. Billy's throw *might have* caused the window to shatter, but as things actually played out, it was Suzy's throw that caused it. Relations of actual causation are important to assessments of moral and legal responsibility.

A natural idea would be to adapt a probabilistic theory of causation of the sort discussed in section 37.2 above. Instead of requiring that a cause C raise the probability of E in *every* background condition B_i, we would require that C raise the probability of E in the background condition that actually occurs.

This natural suggestion runs into a number of problems, which can be grouped in two categories. The first involve cases where actual causes fail to raise the probability of their effects; the second involve cases where non-causes raise the probabilities of their non-effects. I will give one illustration of each kind of case.

Suppose that Billy and Suzy have rocks. Billy is a more accurate thrower than Suzy. If Billy throws his rock, there is a 60% chance that it will hit the window; if Suzy throws, she has a 40% chance of hitting the window. The window will break if either rock hits it. Billy decides that he will give Suzy the opportunity to throw first. In fact, he decides that he will throw his rock only if Suzy doesn't throw hers. (Perhaps a throw by either one of them will set off an alarm, and they will have to run away before getting caught.) Suzy decides to throw her rock; it hits the window and smashes it. By throwing, Suzy lowered the probability that the window would shatter, from 60% to 40%. This is because Suzy's throw *preempts* Billy's throw. Without Suzy's throw, Billy would throw, and have a 60% chance of breaking the glass. But Suzy's throw prevents Billy from throwing, and substitutes her own, less accurate throw. Nonetheless, Suzy's throw caused the window to shatter.

For a case of the second kind, change the example slightly. Suppose that Billy and Suzy throw their rocks simultaneously. As it happens, Suzy's throw hits the window and Billy's misses. Given that Suzy throws her rock, Billy's throw increases the probability that the

[8] See note 2 above.

window would shatter from 40% to 76% (the probability that at least one of them would hit). But Billy's throw did not in fact cause the window to shatter. In the terminology of Schaffer (2001), Billy's throw is a *fizzler*. It had the potential to shatter the window, but it fizzled out, and something else in fact broke the window.

A number of authors have attempted to address one or both of these types of problem, including Lewis (1986), Menzies (1989), Eells (1991, chapter 6), Kvart (1997, 2004), Noordhof (1999), Schaffer (2001), Dowe (2004), Hitchcock (2004a, 2004b), Glynn (2011), and Twardy and Korb (2011). Some of these are hybrid theories, involving probabilities together with counterfactuals or causal processes. I will here present an account that closely follows Glynn's (and is also similar to Kvart's).

Let $\langle \mathbf{V}, \mathbf{G}, P \rangle$ be a causal Bayes net. $\mathbf{V}$ consists of binary variables that are indexed by time. The variable C_t will take the value one or zero depending on whether an event of type C occurs at time t. I will write C_t for $C_t = 1$, and $\sim C_t$ for $C_t = 0$. In addition, I will write $\mathbf{V}_{<t}$ for the set of all variables $X_{t'} \in \mathbf{V}$ with $t' < t$ (and analogously for other inequalities involving time). $\mathbf{G}$ will be a graph representing the causal structure over the variables in $\mathbf{G}$. $\mathbf{G}$ will be subject to the constraint that if X_t is an ancestor of $Y_{t'}$, then $t < t'$. P is a probability distribution that satisfies the Markov and Minimality conditions with respect to $\mathbf{G}$, and assigns positive probability to every combination of values for the variables in $\mathbf{V}$.[9] As we saw in section 3, $\mathbf{G}$ is determined uniquely from P together with information about time order. In addition, for each variable X in $\mathbf{V}$, we have an assignment $@(X)$, which picks out the value of X that actually occurred. For $\mathbf{W} \subseteq \mathbf{V}$, I will abuse notation slightly and write $@(\mathbf{W})$ for the conjunction of propositions of the form $X = @(X)$ where $X \in \mathbf{W}$.

For C_t to be an actual cause of $E_{t'}$, we must have $t < t'$, and $@(C_t) = @(E_{t'}) = 1$. That is, C_t and $E_{t'}$ must both occur, and C_t must precede $E_{t'}$. (The cases where $\sim C_t$ is an actual cause, and where $\sim E_{t'}$ is an actual effect, are exactly parallel.) A necessary condition for C_t to be an actual cause of $E_{t'}$ is that C_t is a *probability-raiser* of $E_{t'}$:

(PR) C_t is a *probability-raiser* of $E_{t'}$ iff $t < t'$, $@(C_t) = @(E_{t'}) = 1$, and there exists $\mathbf{W} \subseteq \mathbf{V}_{\leq t'}$ such that:

$$P(E_{t'} | C_t \,\&\, @(\mathbf{V}_{<\mathbf{t}}) \,\&\, @(\mathbf{W})) > P(E_{t'} | \sim C_t \,\&\, @(\mathbf{V}_{<\mathbf{t}}) \,\&\, @(\mathbf{W})).$$

The idea is that we first hold fixed the actual values of all of the variables earlier than t. This will catch any common causes of C_t and $E_{t'}$ and ensure that the correlation between them is not spurious. In addition, we may also hold fixed the actual values of any further variables no later than t'. If C_t raises the probability of $E_{t'}$ while holding all of these variables fixed, it is a probability-raiser of $E_{t'}$. The further set of variables $\mathbf{W}$ will be called a *revealer* for C_t and $E_{t'}$ since it reveals the probability-raising relation between them.

In our preemption example, we can have as our variables S_t for whether Suzy throws at time t, $B_{t'}$ for whether Billy throws a moment later, and $W_{t''}$ for whether the window shatters at time t''. It is not the case that $P(W_{t''} \mid S_t) > P(W_{t''} \mid \sim S_t)$. As we saw, Suzy's throw lowered the probability that the window would shatter. However, we can also hold fixed

[9] The assumption that all combinations of values receive positive probability is technically convenient. It implies, for example, that all conditional probabilities are well defined. Note, however, it requires a slight modification of our two examples. For instance, in the case of preemption, we must assume that there is some small probability that Billy would throw if Suzy throws first, there is some small probability that the window will shatter even if neither of them throws, etc. This will not change the fundamental structure of the examples.

$\sim B_{t'}$. Holding fixed that Billy doesn't throw, Suzy's throw increases the probability of the window's shattering from near zero to .4. That is, $P(W_{t''} \mid S_t \,\&\, \sim B_{t'}) > P(W_{t''} \mid \sim S_t \,\&\, \sim B_{t'})$. Thus, Suzy's throw is a probability-raiser for the window's shattering.

In addition, an actual cause must not be *neutralized*. This means that there must not be some set of variables that undoes the work of the revealer. Formally:

(N) Let C_t be a probability-raiser for $E_{t'}$. C_t is neutralized relative to $E_{t'}$ if there exists $\mathbf{U} \subseteq \mathbf{V}$ such that:

(i) If $X \in \mathbf{U}$, then either C_t is not a probability-raiser for $@(X)$, or $@(X)$ is not a probability-raiser for $E_{t''}$.
(ii) There is no $\mathbf{W} \subseteq \mathbf{V}_{\leq t'}$ such that

$$P(E_{t'} | C_t \& @(\mathbf{V}_{<\mathbf{t}}) \& @(\mathbf{W}) \& @(\mathbf{U})) > P(E_{t'} | \sim C_t \& @(\mathbf{V}_{<\mathbf{t}}) \& @(\mathbf{W}) \& @(\mathbf{U})).$$

Clause (ii) says that when we hold all of the variables in **U** fixed at their actual values, there is no longer any revealer for C_t and $E_{t'}$. However, we know that if C_t is not a direct cause of $E_{t'}$, the set of all variables that lie on directed paths between C_t and $E_{t'}$ will screen off C_t and $E_{t'}$. So the existence of such a screening-off set by itself does not rule out the possibility of C_t's being an actual cause of $E_{t'}$. Clause (i) rules out the case where the actual value of a screening-off variable X was caused by C_t, and in turn caused $E_{t''}$. X may lie on a directed path from C_t to $E_{t'}$, but the actual value taken by X cannot be part of a chain of probability-raising.

In our example of fizzling, the neutralizing set can be $\{C\}$, where C takes the value 1 if Billy's rock makes contact with the window. In fact, $C = 0$, since Billy's rock did not hit the window. Holding fixed that Billy's rock did not hit the window, Billy's throw did not raise the probability that the window would shatter. Moreover, Billy's throw is not a probability-raiser for $C = 0$. Billy's throw increased the probability that his rock would make contact with the window. Similarly, $C = 0$ is not a probability-raiser for the window's shattering, since $C = 1$ virtually guarantees that the window will shatter. The idea is that Billy's throw is a probability-raiser for the window's shattering because his throw had a tendency to cause his rock to make contact with the window, which in turn had a tendency to cause the window to shatter. However, as the events actually played out, the intermediate event in the chain throw-contact-shatter did not occur. The event that did occur, Billy's rock missing the window, was not one whose possibility helped Billy's throw raise the probability of the window breaking.

Putting the various pieces together, we get the following definition of actual cause:

(AC) C_t is an *actual cause* of $E_{t'}$ iff

(i) $t < t'$,
(ii) $@(C_t) = @(E_{t'}) = 1$
(iii) C_t is a probability-raiser of $E_{t'}$
(iv) C_t is not neutralized with respect to $E_{t'}$

While this kind of account agrees with our ordinary judgments about actual causation in a range of cases, there are still a few potential problems. First, it requires that one start with a fairly rich set of variables. For example, in treating the case of fizzling, it is not enough to have variables for Billy's throw and the window's shattering; one also needs a variable for the intervening event of the rock's making contact with the window. But as with the probabilistic theories of causation discussed in section 37.2, there are no guidelines indicating when we have 'enough' variables. This does not imply that the analysis is wrong, but it may make it impossible in practice to determine whether the analysis yields the intuitively correct verdict.

Secondly, as Hall (2007) and Hitchcock (2007) point out, there are a number of cases that are structurally isomorphic to cases of preemption where we would not judge that one event is an actual cause of the other. Probabilistic versions of these cases will also cause problems for the type of theory described here. Both Hall and Hitchcock propose to resolve the cases by distinguishing between *default* and *deviant* values of variables, and a similar strategy may be helpful in the probabilistic case as well.

Glymour et al. (2010) also raise a number of general problems for theories of actual causation.

37.6 Conclusion

The connections between causation and probability continue to be explored in a wide range of fields, including philosophy, psychology, neuroscience, statistics, epidemiology, econometrics, and computer science. Rapid progress has been made in the last few decades, and this is likely to remain a lively area of research in years to come.

Acknowledgments

Thanks to Frederick Eberhardt, Luke Fenton-Glynn, and Alan Hájek for helpful comments and suggestions.

References

Cartwright, Nancy (1979) Causal Laws and Effective Strategies. *Noûs*. **13**. pp. 419–37.

Cheng, Patricia (1997) From Covariation to Causation: A Causal Power Theory. *Psychological Review*. 104. pp. 367–405

Dowe, Phil (2004) Chance-lowering Causes. In Dowe, Phil and Noordhof, Paul (eds.) *Cause and Chance*. pp. 28–38. London and New York, NY: Routledge.

Dupré, John (1984) Probabilistic Causality Emancipated. In French, Peter, Uehling, Theodore, Jr., and Wettstein, Howard (eds.) *Midwest Studies in Philosophy*. IX. pp. 169–75. Minneapolis, MN: University of Minnesota Press.

Eberhardt, Frederick and Scheines, Richard (2007) Interventions and Causal Inference. *Philosophy of Science*. 74. pp. 981–95.

Eells, Ellery (1991) *Probabilistic Causality*. Cambridge: Cambridge University Press.

Fitelson, Branden and Hitchcock, Christopher (2011) Probabilistic Measures of Causal Strength. In Illari, Phyllis McKay, Russo, Federica, and Williamson, Jon (eds.) *Causality in the Sciences*. pp. 600–27. Oxford: Oxford University Press.

Glymour, Clark, Danks, David, Glymour, Bruce, Eberhardt, Frederick, Ramsey, Joseph, Scheines, Richard, Spirtes, Peter, Teng, Choh Man, and Zhang, Jiji (2010) Actual Causation: a Stone Soup Essay. *Synthese*. 175. pp. 169–92.

Glynn, Luke (2011) A Probabilistic Analysis of Causation. *British Journal for the Philosophy of Science*. 62. pp. 343–92.

Greenland, Sander and Robins, James (1988) Conceptual Problems in the Definition and Interpretation of Attributable Fractions. *American Journal of Epidemiology*. 128. pp. 1185–97.

Hall, Ned (2007) Structural Equations and Causation. *Philosophical Studies*. 132. pp. 109–36.

Herskovits v. Group Health of Puget Sound. 664 P 2d 474. Washington 1983.

Hitchcock, Christopher (2004a) Do All and Only Causes Raise the Probabilities of Effects? In Collins, John, Hall, Ned, and Paul, L. A. (eds.) *Causation and Counterfactuals*. pp. 403–18. Cambridge, MA: MIT Press.

Hitchcock, Christopher (2004b) Routes, Processes, and Chance Lowering Causes. In Dowe, Phil and Noordhof, Paul (eds.) *Cause and Chance*. pp. 138–51. London and New York, NY: Routledge.

Hitchcock, Christopher (2007) Prevention, Preemption, and the Principle of Sufficient Reason. *Philosophical Review*. 116. pp. 495–532.

Hitchcock, Christopher (2010) Probabilistic Causation. In Zalta, E. N. (ed.) *Stanford Encyclopedia of Philosophy*. [Online] Available from: http://plato.stanford.edu/entries/causation-probabilistic/. [Accessed 28 Sep 2015.]

Jeffrey, Richard (1969) Statistical Explanation vs. Statistical Inference. In Rescher, Nicholas (ed.) *Essays in Honor of Carl G. Hempel*. pp. 104–13. Dordrecht: Reidel.

Khoury, Muin, Flanders, W. Dana, Greenland, Sander, and Adams, Melissa (1989) On the Measurement of Susceptibility in Epidemiological Studies. *American Journal of Epidemiology*. 129. pp. 183–90.

Kvart, Igal (1997) Cause and Some Positive Causal Impact. In Tomberlin, James (ed.) *Philosophical Perspectives 11: Mind, Causation, and World*. pp. 401–32. Atascadero: Ridgeview.

Kvart, Igal (2004) Causation: Probabilistic and Counterfactual Analyses. In Collins, John, Hall, Ned, and Paul, L. A. (eds.) (2004) *Causation and Counterfactuals*. pp. 359–87. Cambridge, MA: MIT Press.

Lewis, David (1973/1986) Causation. *Journal of Philosophy*. 70. pp. 556–67. [Reprinted in Lewis, David. *Philosophical Papers*. Volume II. Oxford: Oxford University Press.]

Lewis, David (1979/1986) Counterfactual Dependence and Time's Arrow. *Noûs*. 13. pp. 455–76. [Reprinted in Lewis, David. *Philosophical Papers*. Volume II. Oxford: Oxford University Press.]

Lewis, David (1986) Chancy Causation. In *Philosophical Papers*. Volume II. pp. 175–84. Oxford: Oxford University Press.

Menzies, Peter (1989) Probabilistic Causation and Causal Processes: A Critique of Lewis. *Philosophy of Science*. 56. pp. 642–63.

Noordhof, Paul (1999) Probabilistic Causation, Preemption and Counterfactuals. *Mind*. 108. pp. 95–125.

Parascandola, Mark (1996) Evidence and Association: Epistemic Confusion in Toxic Tort Law. *Philosophy of Science*. 63. pp. S168–S176.

Pearl, Judea (1988) *Probabilistic Reasoning in Intelligent Systems.* San Francisco, CA: Morgan Kaufman.

Pearl, Judea (2009) *Causality: Models, Reasoning, and Inference*. 2nd edition. Cambridge: Cambridge University Press.

Pearl, Judea and Verma, Thomas (1991) A Theory of Inferred Causation. *Principles of Knowledge Representation and Reasoning: Proceedings of the 2nd International Conference.* pp. 441–52. San Mateo, CA: Morgan Kaufman.

Reichenbach, Hans (1956) *The Direction of Time*. Berkeley and Los Angeles, CA: University of California Press.

Rothman, Kenneth, Greenland, Sander, and Ash, Timothy (2008) *Modern Epidemiology*. 3rd edition. Philadelphia, PA: Lippincott Williams & Wilkins.

Schaffer, Jonathan (2001) Causes as Probability-Raisers of Processes. *Journal of Philosophy*. 98. pp. 75–92.

Sheps, Mindel (1958) Shall we Count the Living or the Dead? *New England Journal of Medicine*. 259. 12. pp. 210–14.

Skyrms, Brian (1980) *Causal Necessity*. New Haven, CT and London: Yale University Press.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard (2000) *Causation, Prediction and Search*. 2nd edition. Cambridge, MA: MIT Press.

Suppes, Patrick (1970) *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.

Twardy, Charles, and Korb, Kevin (2011) Actual Causation by Probabilistic Active Paths. *Philosophy of Science*. 78. pp. 900–13.

Woodward, James (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.

Wright, Sewall (1921) Correlation and Causation. *Journal of Agricultural Research*. 202. pp. 557–85.

Name Index

Abadi, M. 687, 699
Abdallah, F. 479–99
Abrams, Marshall 592, 597, 606, 608, 619
Abramsky, S. 555, 565, 570
Aczél, J. 524, 525, 539
Adams, Ernest W. 168, 181, 182, 300, 308, 681, 682, 684, 687, 689, 690, 691, 692, 693, 696, 699, 774, 776, 777, 779, 780, 785
Adams, Melissa 831
Adolphs, Ralph 302, 310
Agnoli, F. 486, 490
Airy, G. B. 116, 126
Albert, David Z. 436, 437, 467, 473, 546, 561, 570, 588, 593, 596, 597
Albrecht, A. 584, 593, 597
Aldrich, John 2, 113, 119, 120, 122, 123, 126, 127, 386, 403
Allais, Maurice 367, 379, 726, 734, 810, 811, 812
Aliotta, Antonio 142
Alpert, M. 484, 490, 509, 515
Anderson, B. D. O. 301, 311, 515, 662
Angner, Eric 718, 734
Appiah, Kwame Anthony 630, 644, 776, 785
Aquinas, Thomas 41, 47
Arbuthnot, John 70, 72, 74, 75
Ariew, Andre 608, 613, 619, 620
Aristotle 34–8, 42
Arló-Costa, Horacio 688, 689, 692, 699
Armendt, Brad 363, 379, 628, 644, 794, 812
Arntzenius, Frank 180, 181, 430, 437, 632, 636, 644, 678, 679
Arrow, Kenneth J. 537, 539, 540
Ash, Timothy 832
Augustin, T. 298, 302, 308, 309, 312, 323, 337
Aumann, Robert. J. 300, 308
Austin, J. L. 381, 772, 773, 785
Avicenna (Ibn Sina) 41, 42
Ayer, Alfred Jules 409, 420
Azo of Bologna 35, 47

Bacchus, F. 686, 700
Bacciagaluppi, Guido 4, 546, 565, 570
Bacharach, M. 380, 523, 539, 813
Bachelier, Louis 107, 109
Bailey, J. 570
Balasubramanian, V. 439
Balding, D. 209, 244
Baldus de Ubaldis 36
Baltag, A. 689, 700
Bamber, D. 692, 700
Barasz, M. 700
Bar-Hillel, M. 482, 490, 507, 515, 571
Barlas, S. 493, 517
Barnett, J. A. 300, 308
Barnett, Vic 358, 359
Baron, J. 490, 494
Barron, G. 480, 491, 492
Barrow, J. D. 641, 644, 741, 768
Bartha, Paul 4, 159, 162, 165, 762, 763, 765, 768, 792, 812
Bassi, A. 593, 594, 597
Bavli, G. M. 108
Bayes, Thomas 70, 82–3, 84, 86, 88, 89, 91, 316, 322, 323, 326, 336, 337, 338, 340, 738, 768, 769
Beatty, John H. 604–616, 618, 619, 620, 621
Beck, J. R. 485, 495
Beebee, Helen 435, 437
Bell, John S. 548, 561, 563, 564, 565, 568, 569, 570, 571

Bellhouse, David R. 2, 45, 46, 47, 55, 66, 69, 70, 74, 75, 76, 77, 78, 80, 81, 82, 83, 84, 85
Beltrametti, E. G. 555, 570
Benferhat, S. 694, 697, 700
Ben-Haim, Yakov 302, 308
Bennett, J. H. 127, 691, 699, 700, 782, 786
Berger, J.O. 300, 302, 303, 308, 384, 393, 394, 395, 400, 402, 403
Berger, R. L. 393, 399, 403
Berkovitz, J. 454, 457, 463, 473, 590, 597
Bermúdez, José 795, 796, 807, 810, 812
Bernardo, J. M. 381, 384, 394, 395, 402, 403, 516
Bernoulli, Daniel 66, 727, 790, 812
Bernoulli, Jacob (also 'Jakob', 'James') 2, 50–68, 69, 71, 73, 74, 77, 78, 85, 237, 298, 311, 316, 317–8, 323, 337, 338, 346, 347, 477
Bernoulli, Johann 77
Bernoulli, Nicolaus I (also 'Nicholas', 'Nicolas'; nephew of Jacob) 53, 54, 68, 339, 790
Bernoulli, Nicolaus (son of Jacob) 54, 61, 67
Bernshtein, Sergei Natanovich 102, 105, 106, 109
Bernstein, A. R. 428, 437, 631, 644
Bernstein, S. N. 299, 308
Berry, Andrew C. 103, 400, 403
Bertrand, Joseph L. F. 16, 29, 92, 179, 181, 203, 334, 335, 337
Bessel, Friedrich Wilhelm 90
Bewley, T. F. 301, 308
Beyth-Marom, R. 487, 488, 490
Bhatt, M. 302, 310
Bienaymé, Irenée Jules 93, 94, 103, 104, 109
Bigelow, John 437, 466, 473
Billingsley, Patrick 240, 244
Birkhoff, G. D. 555, 570, 590, 597
Birnbaum, A. 384, 385, 403
Black, Max 212, 218, 244
Black, R. 434, 437
Blackburn, Simon 777, 786
Blackorby, C. 720, 734
Blais, A-R. 490, 496
Blum, J. R. 300, 307, 308
Boardley, R. F. 512, 515
Boettcher, W.A. 488, 490
Bohlmann, Georg 99, 109
Bohm, David 431, 436, 467, 568, 571
Bolker, Ethan 269, 273, 792, 812
Boltzmann, Ludwig 95–7, 106, 111, 573, 574, 576, 578, 581–4, 587–90, 597, 598, 599
Bolzano, Bernard 130
Bonano, E. J. 500, 515
Bonini, Alexander 44, 47
Boole, George 126, 131, 298, 308, 319, 323, 335, 337, 340, 564, 570, 571, 681, 700
Borel, Émile 98, 99, 103, 104, 109, 133, 351, 447, 448, 455, 457, 470, 477, 478, 479, 489, 490, 491
Bossert, W. 720, 734
Bostrom, Nick 669, 678, 679, 749, 768
Bouchard, Frederic 606, 608, 619
Bovens, Luc 142, 148
Box, J. F. 122, 126
Bradley, Darren 642, 644, 669, 673, 678, 679
Bradley, Richard 527, 537, 539
Brandenburger, Adam 565, 566, 570
Brandon, Robert N. 604, 605, 607, 608, 609, 615, 616, 619, 620
Breeding, R. J. 517
Brenner, L. A. 509, 514
Brier, Glenn W. 143, 144, 261, 371, 379, 489, 491, 506, 515
Broad, C. D. 125, 326, 337
Brodén, Torsten 99
Broglie, Louis de 436, 568
Bronfman, Aaron 630, 644
Brooks, R. A. 66, 188, 199
Broome, John 720, 721, 724, 725, 727, 734, 803, 804, 805, 812
Broomell, S. B. 244, 491, 495
Brown, H. R. 578, 598
Brown, L. D. 402, 403
Browne, William 80, 86
Brun, W. 211, 244, 487, 488, 491, 495
Brush, Stephen G. 108, 109, 573, 598
Buchak, Lara 4, 26, 683, 700, 713, 734, 799, 810, 812
Buchanan, B. G. 188, 199
Buck, C. E. 494, 515

Budescu, D. V. 211, 212, 224, 244, 487, 488, 489, 491, 495, 496
Buehler, R. J. 301, 308
Bulmer, M. G. 113, 114, 127
Burgess, John P. 687, 700
Burgin, M. 163, 165, 166
Burian, Richard M. 604, 606, 607, 620
Burmaster, D. E. 503, 516
Busch, P. 547, 554, 570
Butterfield, Jeremy 463, 473, 474, 600

Cabantous, L. 479, 491
Cahan, A. 210, 244
Caie, Michael 274, 687, 700
Calderoni, Mario 142
Callender, Craig 426, 427, 437
Callinger, R. 65, 67
Camerer, Colin F. 302, 310, 479, 480, 491
Cameron, Robert 109
Camilleri, A. R. 481, 491
Campolongo, F. 518
Cannon, C. M. 300, 308
Cano, M. 40, 47
Cantelli, Francesco 104, 110
Canton, John 82, 84, 86
Caraco, T. 490, 491
Caramuel Lobkowitz, J. 38
Cardano, Giralamo 36, 46, 47, 53, 56, 57, 66, 67
Carlson, B. W. 486, 496
Carnap, Rudolph 131, 132, 135, 138, 145, 148, 151, 316, 324, 326, 327, 329, 331, 337, 340, 360, 384, 403, 630, 639, 640, 644, 654, 656, 659, 681, 700
Carroll, Sean 588, 597, 598
Carter, Brandon 743, 768
Cartwright, Nancy 814, 818, 830
Carvalhaes, C. 555, 571
Casella, G. 393, 399, 403
Casscells, W. 487, 491
Cassinelli, G. 555, 570
Cattaneo, G. 552, 570
Cauchy, Augustin Louis 91, 93, 110
Caves, Carlton 215, 244
Ceccarelli, G. 44, 47
Chaitin, Gregory 457, 449, 450, 451
Chalmers, David 20, 29, 667, 678
Chambers, C. 525, 539
Chandler, Jake 644, 647, 768, 769
Chang, C. C. 696, 700
Chebyshev, Pafnutii Lvovich 91, 93, 94, 95, 98, 101, 103, 104, 105, 110
Cheng, Patricia 826, 830
Chierchia, G. 776, 786
Choquet, G. 259, 274, 298, 300, 301, 302, 303, 304
Christakopoulou, S. 483, 493
Christensen, David 175, 181, 246, 626, 628, 632, 635, 645, 682, 700, 796, 799, 812
Christiano, P. 687, 700
Chung, Kai Lai 20, 29
Church, Alonzo 345, 359, 446, 447, 448, 449, 451, 455, 457
Cicero 34, 35, 48, 49, 69
Clark, P. 149, 455, 458, 463, 473, 831
Clauser, J. F. 563, 570
Clausius, Rudolph 96, 576, 578, 579, 598
Clemen, R. T. 482, 492, 501, 508, 512, 515, 516
Cohen, B. L. 487, 492, 494
Cohen, Jacob 392, 403
Cohen, John 478, 491
Cohen, Jonathan 426, 427, 437
Cohen, L. J. 487, 491
Cohen, Robert S. 381, 571, 648
Coletti, G. 301, 309
Collins, John 437, 466, 473, 831
Collins, Robin 641, 645, 741, 743, 745, 748, 749, 769
Colyvan, Mark 601, 620, 743, 744, 747, 748, 749, 769, 792, 812
Conder, M. 231, 244
Condorcet, Marie Jean Antoine Nicolas Caritat, Marquis de 88
Cook, Alan 138, 148
Cooke, R. M. 501, 511, 513, 514, 515, 516
Coolen, F. P. A. 302, 308, 309, 494
Coolen-Schrijner, P. 302, 309
Cooper, G. F. 188, 192, 199
Copernicus, Nicolaus 39
Corani, G. 302, 309
Cosmides, L. 487, 491, 493, 494
Coumet, E. 45, 48

Cournot, Antoine Augustin 92, 315, 323, 336, 337
Cox, D. R. 346, 359
Cox, Richard 139
Cozman, Fabio G. 2, 3, 172, 232, 243, 278, 294, 302, 307, 309, 361, 488, 492, 640, 645
Craig, John 70, 73, 74
Craig, William L. 743, 769
Cramér, Harald 101, 102, 109, 110
Cross, Charles B. 684, 700
Crupi, Vincenzo 3, 4, 175, 490, 492, 650, 651, 652, 654, 656, 658, 659, 660, 662, 740, 769
Cumberland, Richard 70, 72, 85
Cuming, Sir Alexander 79, 80
Cumming, G. 396, 403

Dale, Andrew I. 322, 337, 339
Dalkey, N. 510, 512, 515
Daneshkhah, A. 494, 515
Danks, David, 830, 831
Darby, G. 463, 473
Darwin, Charles 115, 116, 120, 123, 127, 128, 374, 375, 602, 603, 604, 605, 606, 612, 615
Dasgupta, A. 444, 448, 458
Daston, Lorraine J. 55, 67, 318, 337
Davidson, Donald 801, 812
Davies, Martin 643, 645
Dawid, Philip A. 144, 148
de Barros, J. A. 555, 571
De Blic, J. 37, 48
De Cooman, G. 244, 298, 302, 305, 307, 308, 309, 312, 494
de Finetti, Bruno 3, 103, 130, 132, 138, 139, 142–7, 148, 149, 159, 162, 170, 179, 181, 182, 206, 224, 225, 245, 261, 262, 274, 275, 277, 294, 299, 300, 301, 305, 309, 316, 319, 324, 325, 328, 329, 330, 336, 337, 356, 359, 360, 362, 363, 371, 379, 478, 492, 502, 515, 574, 628, 645, 689, 787, 799, 812
de Lugo, J. 41, 45, 48
De Moivre, Abraham 53, 54, 61, 64–6, 67, 69–72, 74–84, 85, 86, 107, 316, 317, 337
De Morgan, Augustus 83, 84, 90, 131, 317, 318, 337, 340, 477
de Witt, Jan 36, 62
Delbecq, A. L. 510, 516
Dembski, William 743, 769, 770
Diamond, P. 721, 722, 735
Dietrich, Franz 3, 4, 526, 527, 530, 531, 532, 535, 536, 537, 539, 540
Djulbegovic, Benjamin 601, 620
Dodd, N. G. 503, 508, 516
Doeblin, Wolfgang 103
Dokow, E. 536, 540
Domingos, P. 683, 703
Donaldson, Donald 710, 734
Donkin, William Fishburn 133, 477
Donsker, Monroe D. 109
Doob, Joseph Leo 109, 110
Dorr, Cian 636, 645
Dowe, Phil 828, 830, 831
Dreier, James 795, 805, 812
Drouet, Isabelle 618, 620
Druzdel, M. J. 502, 516
Dubois, D. 228, 274, 700
Duddy, C. 537, 540
Duff, Anthony 758, 759, 764, 767, 769
Dummett, Michael 248, 256, 274, 275, 565, 570
Duns Scotus 41, 48
Dupré, John 620, 818, 830
Dyer, J. S. 503, 517
Dzhafarov, E. 566, 571

Easwaran, Kenny 2, 17, 18, 181, 282, 294, 628, 645, 667, 764, 765, 766, 767, 769, 792, 812
Eberhardt, Frederick 138, 149, 823, 830, 831
Eddington, Arthur 749, 751, 769
Edgeworth, F. 113, 116, 118, 323, 715
Edgington, Dorothy 169, 181, 692, 693, 699, 700, 775, 776, 777, 778, 779, 780, 781, 786
Edwards, A. W. F. 57, 67, 322, 338, 386, 400, 403, 655, 660, 741, 769
Edwards, P. 244
Edwards, Ward 403, 478, 481, 492, 494, 514, 515, 630, 645, 735

Eells, Ellery 182, 375, 376, 379, 658, 660, 786, 815, 818, 828, 831
Egan, Andy 785, 786, 787, 788
Ehrenfest, Paul 96, 98, 110, 576, 581, 582, 598
Ehrenfest, Tatjana 96, 98, 110, 576, 581, 582, 598
Einstein, Albert 2, 108, 110, 326, 371, 374, 375
Elga, Adam 180, 181, 428, 438, 457, 458, 473, 632, 641, 645, 677, 679, 812
Ellis, Brian 779, 786
Ellis, G. 641, 645
Ellis, Robert Leslie 132, 319, 323
Ellsberg, Daniel 297, 299, 301, 307, 309, 479, 492, 811, 812
Ellsberg, M. 726, 735
Engel, Eduardo M. R. A. 321, 338, 598
Epley, N. 481, 492
Erev, I. 480, 487, 492, 493, 494, 487, 493, 496
Eriksson, Lina 214, 245, 369, 370, 379
Ernst, Zachary 619, 620
Esparza, M. de 41
Esseen, Carl Gustav 101
Essig, A. 245
Etchart-Vincent, N. 492, 480
Evans, J. S. 489,492, 612, 669, 679
Everitt, C. W. F. 598

Fagin, R. 683, 686, 687, 700
Fairley, W. 209, 245
Falk, R. 334, 338
Feigenbaum, E. A. 199
Feller, William (also 'Willy', 'Vilim') 102, 103, 110, 318, 326, 338, 349, 359
Fenstad, J. E. 683, 700
Fermat, Pierre 2, 33, 36, 45, 46, 53, 54, 56, 57, 65, 316, 373, 812
Fetzer, James H. 413, 415, 416, 417, 418, 321
Fidler, Fiona 392, 396, 403
Fiedler, K. 486, 492
Field, Hartry 248, 251, 254, 255, 256, 270, 275, 279, 294, 684, 693, 700
Fierens, P. I. 298, 309
Filon, L. N. G. 119, 123, 129
Finch, S. 396, 403
Fine, Arthur 564, 571
Fine, Kit 251, 275
Fine, Terrence L. 2, 17, 27, 163, 164, 165, 218, 222, 224, 225, 226, 231, 235, 238, 239, 241, 242, 245, 246, 247, 298, 299, 301, 303, 309, 310, 311, 312, 341, 346, 347, 348, 350, 354, 359
Finkelstein, M. 209, 245
Finsen, Susan (also Susan Mills) 604–612, 614, 615, 616, 618, 619, 620
Fioretti, G. 298, 309
Fischer, Hans 2, 27, 93, 101, 103, 106, 108, 109, 110
Fischhoff, Baruch 30, 482, 484, 492, 493, 505, 508, 509, 516, 517
Fishburn, Peter C. 299, 300, 306, 307, 309, 367, 368, 369, 379, 725, 735, 812
Fisher, Ronald A. 112, 113, 114, 117, 119, 121, 122–4, 125, 126, 127, 129, 139, 382, 387, 388–90, 391, 392, 395, 397, 398, 399, 403
Fisk, J. E. 486, 492
Fitelson, Branden 30, 490, 492, 651, 652, 655, 656, 657, 658, 660, 827, 831
Flanders, W. Dana 831
Fleming, J. 38, 48
Fodor, Jerry 468, 473
Fokker, Adriaan Daniel 108
Foley, Richard 626, 645, 683, 700
Fontenelle, Bernard Le Bouyer de 53
Foot, Philippa 711, 735
Forrest, Peter 782, 786
Forsyth, B. H. 496
Foster, D. P. 484, 485, 496
Foulis, D. J. 555, 571, 572
Fox Keller, Evelyn 613, 620
Fox, Craig R. 480, 482, 483, 492, 508, 510, 516
Fox, Danny 776, 786
Frankish, K. 626, 645
Franklin, James 2, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 59, 67
Fraser, D. A. S. 180, 335, 338
Fréchet, Maurice 100, 105, 109, 110, 470
Freedman, D. 329, 330, 337
French, Steven 511, 516
Friedberg, A. 48
Fries, Jakob Friedrich 92, 323

Friesen, J. 72, 85
Frigg, Roman 4, 437, 438, 454, 457, 468, 469, 470, 473, 597, 598
Frisch, Mathias 437, 438, 686, 693, 701
Fristedt, B. 241, 245
Fuchs, Christopher 244
Fuhrmann, Andre 697, 701

Gabbay, Dov M. 43, 48, 149, 405, 570, 648, 651, 660, 697, 701, 702, 703
Gaifman, Haim 428, 438, 447, 458, 683, 686, 701
Galavotti, Maria Carla 2, 3, 148, 149, 337, 473, 598, 660
Gale, Richard 764, 769
Galen 40, 48
Galileo 39, 46, 48
Galton, Francis 112, 113, 114–7, 118, 119, 120, 121, 122, 123, 126, 127, 128, 129, 386
Garber, Daniel 376, 379, 577
Garber, E. 577, 598
Gardenfors, Peter 301, 306, 310
Gardner, Martin 16, 30
Garfield, Jay 769
Garthwaite, P. H. 485, 492, 494, 515
Gauss, Carl Friedrich 89, 90, 110, 113, 206, 329
Gauthier, David 732, 735
Gayon, Jean 113, 128, 604, 620
Gelfand, M. J. 483, 493
Gelman, A. 402, 403
Genest, C. 70, 85, 511, 516, 521, 526, 529, 540
Georgescu, G. 274
Gerbrandy, J. 687, 704
Gerla, B. 258, 275
Gerson, J. 36, 37
Ghirardato, Paolo 160, 165
Ghirardi, G. C. 436, 438, 473, 571, 593, 594, 597, 598
Ghurye, S. G. 312
Gibbard, Allan 778, 779, 780, 784, 786, 793, 812
Gibbs, Josiah Willard 96, 97, 573, 574, 577, 580, 585–8, 589, 590, 593, 598, 599
Giere, Ronald N. 412, 413, 414, 422, 424, 438
Gigerenzer, Gerd 393, 404, 479, 482, 489, 490, 493, 509, 516, 620
Gilboa, Itzhak 301, 304, 306, 307, 308, 309, 310, 707, 713, 724, 726, 735, 799, 812
Gilio, A. 700
Gill, J. 231, 245
Gillespie, John H. 609, 610, 620
Gillham, N. W. 113, 128
Gillies, Anthony 775, 786
Gillies, Donald 4, 134, 149, 323, 338, 389, 390, 404, 420, 422, 440, 611, 620, 740, 769
Gilon, D. 244
Gilovich, Thomas 481, 492, 507, 515, 518
Giron, F. J. 301, 306, 307, 310
Gleason, A. M. 554, 565, 571
Glymour, Bruce 830, 831
Glymour, Clark 16, 30, 138,149, 199, 374, 379, 650, 661, 819, 821, 830, 831, 832
Glynn, Luke 455, 458, 468, 470, 473, 828, 830, 831
Gnedenko, Boris 103, 110, Vladimirovich
Gneiting, T. 517
Goel, P. K. 182
Goldblatt, R. 756, 769
Goldstein, M. 302, 310
Goldstein, Sheldon 582, 583, 598
Goldszmidt, M. 692, 701
Gonzales, R. 480, 493
Gonzalez-Vallejo, C. C. 493, 517, 484, 489
Good, Irving John 131, 148, 149, 299, 300, 301, 310, 315, 327, 338, 653, 654, 656, 660
Goodman, Nelson 325, 338, 656
Goodman, Noah 215, 247
Goodman, S. N. 390, 404
Goossens, L.H. J. 516
Gordon, B. 43, 49
Gorry, G. A. 210, 245
Gosset, William Sealy (also known as 'Student') 120–4, 129, 386
Gower, B. 335, 338
Grabowski, M. 547, 570
Graboys, T. B. 491
Graunt, J. 39

Gray, L. 241, 245
Greaves, Hilary 635, 645
Greenhalgh, Trisha 209, 245
Greenland, Sander 342, 359, 601, 620, 826, 831, 832
Gregory, David 74, 85
Grice, Paul 775, 776, 779, 786
Griffiths, T. L. 215, 247
Guerrini, A. 74, 85
Gurr, M. 495
Gustafson, D. H. 510, 516
Guzowski, R.V. 501, 518

Haack, Susan 74, 275
Hacking, Ian 2, 33, 38, 48, 61, 65, 67, 318, 319, 321, 338, 386, 389, 404, 420, 421, 422, 458, 574, 598, 660, 738, 769
Hadamard, Jacques 108
Hadar, L. 480, 492
Haddawy, P. 686, 693, 701
Haenni, R. 681, 701
Hagen, Gotthilf Heinrich Ludwig 90
Hagen, O. 379, 811, 812
Hailperin, T. 298, 300, 310, 685, 688, 701
Haim, E. 507, 516
Hájek, Alan 30, 33, 48, 160, 161, 162, 163, 165, 169, 170–3, 174, 175, 182, 205, 214, 215, 224, 233, 243, 244, 245, 249, 262, 265, 267, 275, 294, 309, 310, 312, 341, 342, 346, 348, 350–8, 359, 361, 363, 364, 365, 369. 370, 372, 379, 380, 403, 405, 406, 413, 421, 422, 429, 438, 451, 453, 457, 458, 461, 464, 472, 473, 474, 492, 537, 575, 598, 616, 620, 627, 629, 630, 631, 634, 645, 647, 651, 655, 657, 660, 691, 699, 701, 707, 710, 734, 735, 737, 740, 758–5, 767, 769, 770, 771, 777, 786, 792, 813, 830
Hald, Anders 61, 64, 65, 67, 70, 77, 83, 85, 88, 110,112, 113, 124, 128, 322, 338
Hall, Ned 169, 182, 430, 433, 434, 437, 438, 633, 645, 777, 786, 830, 831
Halley, Edmond 81, 82, 84
Halpern, Joseph Y. 174, 182, 243, 245, 672, 679, 684, 686, 687, 688, 692, 693, 699, 700, 701, 769
Halpin, J. 633, 645
Hamblin, C. L. 218, 245, 687, 701
Hamm, R. M. 487, 493
Hammond, P. 718, 724, 735, 737, 769, 797
Hansen, Lars Peter 301, 310
Hansen, Pierre 301, 310
Hansson, B. 813
Hansson, P. 485, 493, 496, 508, 518
Harlow, L. L. 392, 404
Harman, Gilbert 158, 165, 275
Harper, F. T. 500, 516
Harper, William L. 338, 340, 380, 381, 704, 786, 787, 793, 812, 814
Harris, J. 76, 85
Harsanyi, John 706, 716, 717, 718, 719, 720, 721, 722, 723, 724, 725, 726, 727, 728, 729, 731, 732, 733, 735
Harting, John 614, 621
Hartmann, Stephan 148, 149, 392, 404
Harvey, W. 72, 85
Hatori, T. 495
Haug, E. G. 163, 165
Hausner, M. 724, 735
Hawthorne, James 2, 173, 180, 181, 224, 281, 295, 296, 299, 310, 458, 657, 658, 659, 660, 661, 679, 683, 688, 689, 701, 814
Heckerman, D. 185, 188, 199, 661
Heidelberger, M. 321, 338
Heifetz, A. 686, 702
Hell, W. 494
Hellman, Geoffrey 451, 453, 456, 458
Helton, J. C. 503, 516
Hempel, Carl G. 652, 653, 656, 661, 683, 702
Henrion, M. 515, 517
Hermann, Jacob 53, 151, 337
Herreshoff, M. 700
Herskovits, E. 192, 199, 825, 831
Hertwig, R. 480, 493
Herzberg, F. 537, 540
Heyde, C. C. 113, 128
Hilbert, David 97, 98, 99, 110
Hill, Bruce M. 178, 180, 182, 336, 338
Hilpinen, R. 683, 702
Hilton, D. J. 488, 496
Hintikka, Jaakko 149, 245, 246, 312, 328, 338, 662, 699, 703, 704, 735
Hirschfeldt, D. R. 450, 458

Hitchcock, Christopher 162, 165, 205, 224, 243, 245, 294, 309, 310, 645, 647, 679, 827, 828, 830, 831
Ho, T. H. 480, 491
Hobbes, Thomas 42, 48, 70, 72, 85
Hochkirchen, Thomas 100, 104, 110
Hodge, M. J. S. 606, 616, 619, 620
Hodges, J. L., Jr. 300, 310
Hoefer, Carl 429, 431, 433, 434, 437, 438, 463, 464, 465, 468, 469, 470, 472, 473, 708, 735
Hoeffding, W. 312
Höflich, Paul 108
Hoffrage, U. 509, 516
Hogarth, R. M. 478, 493
Hogger, C. J. 701
Holder, R. 641, 645
Holt, R. A. 563, 570
Holzman, R. 537, 540
Hooker, Clifford A. 570, 571
Hoover, D. N. 686, 702
Hoover, Kevin D. 393, 404
Hopf, E. 321, 470, 598, 599
Hoppe, F. 338
Hora, J. A. 516, 517
Hora, S. C. 3, 4, 501, 503, 505, 508, 512, 514, 515, 516, 517, 518
Horne, M. A. 570
Horowitz, Tamara 369, 380
Hostinský, Bohuslaw 108
Hotelling, H. 112, 125, 126, 128, 312
Howie, David 149
Howson, Colin 149, 158, 166, 275, 335, 380, 404, 410, 411, 422, 645, 656, 681, 702, 769
Hozo, Iztok 601, 620
Hsee, C. K. 489, 494
Hsu, M. 302, 310
Huber, Franz 275, 379, 380, 438, 645, 646, 657, 658, 661, 683, 702
Huber, P. J. 300, 303, 304, 310
Hughes, R. I. G. 270, 275
Hume, David 40, 41, 83, 145, 321, 323, 326, 328, 416, 431, 434, 436, 452, 760, 738, 767, 797, 813
Humphreys, Paul W. 148, 149, 150, 155, 161, 166, 169–71, 182, 245, 415–7, 422, 455, 458
Hunter, D. 626, 645
Hurd, A. E. 756, 770
Hurley, Susan 380, 804, 813
Hurwicz, L. 299, 307, 310
Hutter, Marcus 440, 459
Huygens, Christiaan 36, 50, 52–9, 63, 65, 67, 69–76, 78, 79, 80, 84, 85, 86

Iglehart, J. K. 209, 245
Iman, R. L. 501, 514, 517
Innocent III, Pope 35–6
Insua, D. R. 300, 302, 304, 306, 310
Ioannides, John P. A. 395, 404
Ismael, Jenann T. 463, 473

Jackson, E. A. 588, 599
Jackson, Frank 438, 630, 645, 709, 710, 731, 734, 735, 776, 786
Jackson, John P., Jr. 604, 619, 620
Jaffray, J-Y. 258, 275
Jansana, R. 270, 275
Jaumard, B. 301, 310
Jaynes, E. T. 34, 49, 139, 155, 159, 160, 166, 246, 334, 335, 338, 599, 681
Jeffrey, Richard C. 30, 132, 142, 146, 148, 150, 151, 166, 233, 246, 268, 269, 272, 273, 275, 276, 332, 333, 337, 338, 350,357, 359, 361, 363, 364, 376–8, 380, 438, 451, 458, 474, 575, 599, 627, 628, 630, 645, 646, 736, 758, 759, 763, 770, 773, 779, 786, 792, 793, 813, 816, 831
Jeffreys, Harold 3, 119, 125, 128, 130, 131, 132, 138, 139, 140, 141, 142, 144, 145, 146, 147, 148, 150, 151, 335, 339, 384, 403, 404, 630, 646
Jenkinson, D. J. 494, 515
Jensen, K. 724, 735
Jevons, W. Stanley 131, 325, 339
Jiang, Xia 2, 188, 199, 820
Johnson, M. 699, 702
Johnson, R. J. 488, 495

Johnson, William Ernest 125, 131, 138, 151, 316, 323, 324, 326, 327, 331, 339, 340
Jones, J. A. 516
Jonsen, A. R. 36, 48
Joyce, James M. 175, 182, 235, 246, 261, 275, 369, 371, 372, 380, 430, 438, 629, 630, 646, 651, 657, 660, 732, 735, 786, 794, 814
Jungermann, H. 486, 495
Juslin, P. 485, 493, 496, 508, 509, 517, 518
Justinian, Emperor 48

Kadane, Joseph B. 166, 182, 311, 379, 381, 401, 404, 489, 494, 680, 770
Kahle, Ludwig Martin 69, 86
Kahneman, Daniel 367, 380, 478, 479, 480, 481, 482, 485, 486, 487, 490, 492, 493, 495, 496, 507, 508, 515, 517, 518, 806, 807, 813
Kamm, Frances 733, 735
Kant, Immanuel 416, 422
Kantola, I. 37, 48
Kaplan, David 669, 679
Kaplan, Jonathan 608, 621
Kaplan, M. A. 224, 225, 226, 246, 299, 310, 626, 646
Kaplan, S. 511, 517
Kass, R. 385, 404
Kassirer, J. 210, 245, 246
Katz, Bernard D. 335, 339
Kaufmann, Stefan 701, 703, 784, 786
Kavka, Gregory 732, 735
Keefe, Rosanna 275
Keeney, R. L. 30, 513, 515, 517
Keisler, H. Jerome 299, 310, 686, 696, 700, 702
Keller, J. B. 470, 473, 613, 620
Kelly, Kevin T. 661, 683, 703
Kemeny, John 380, 646
Kemp, C. 215, 247
Kent, S. 488, 493, 502, 517
Kepler, Johannes 39
Keren, G. 509, 517
Kesselman, R. F. 488, 493
Keynes, John Maynard 34, 38, 49, 131, 138, 140, 160, 166, 224, 246, 278, 282, 295, 298, 310, 319, 320, 323, 325, 339, 360, 477, 493, 695, 661, 681
Khinchin, Aleksandr Yakovlevich 103, 104, 105, 108, 110, 111, 470, 591, 599
Khoury, Muin 826, 831
Kirk, G. S. 317, 339
Kitcher, Philip 606, 620
Klayman, J. 484, 485, 493, 495, 509, 517
Kleinbolting, H. 509, 516
Kleinmuntz, D. N. 503, 517
Kleiter, G. D. 697, 703
Kmietowicz, Z. W. 300, 309
Knebel, S. 42, 48
Knight, Frank H. 299, 301, 307, 310, 503, 517
Knott, C. G. 579, 599
Kochen, Simon 270, 555, 561, 563, 564, 565, 569, 571
Koehler, D. J. 482, 486, 496, 509, 510, 515, 518
Kölbel, Max 781, 786
Koller, D. 700
Kolmogorov, Andrei Nikolaevich 2, 16–9, 20, 30, 52, 98, 100, 103, 104, 105, 107, 108, 109, 110, 111, 155–66, 172, 173, 174, 177–81, 182, 205, 206, 214, 238, 240, 246, 247, 269, 345, 351, 356, 359, 417, 422, 442, 449, 450, 451, 458, 627, 646
Kolodny, Niko 800, 813
Kooi, B. P. 687, 702, 704
Koopman, Bernard O. 278, 295, 299, 310
Koppel, M. 43, 48
Korb, Kevin 828, 832
Kotzen, Matthew 3, 4, 361, 644, 646, 770
Kraan, B, 516
Kraft, C. 224, 227, 231, 246, 275
Krantz, D. H. 215, 246, 486, 490, 727, 736
Kratzer, Angelika 516, 778, 786
Kraus, S. 683, 694, 697, 698, 702
Krauss, P. 703
Kripke, Saul 435, 438, 784
Krönig August Carl 95, 576, 599
Kronz, Frederick 597
Kruger, L. 386, 404
Kruse, M. 400, 404
Kuhn, Thomas 380

Kuipers, Theo A. F. 328, 339, 650, 661
Kujala, J. V. 571
Kumar, A. 224, 246
Kuznetsov, V. P. 302, 310
Kvart, Igal 828, 831
Kyburg, Henry E. Jr. 3, 163, 166, 188, 234, 246, 294, 299, 306, 310, 311, 479, 491, 493, 626, 637, 638, 639, 636, 65, 661, 683, 688, 702, 704

La Caze, Adam 3, 213, 405, 424, 440, 442, 461, 474
Lahti, P. J. 547, 570
Lando, T. 688, 702
Lane, David A. 179, 182, 336, 339
Lange, Marc 108, 435, 438
Langevin, Paul 108
Langford, C. H. 786
Laplace, Pierre Simon, Marquis de 77, 83, 84, 87–95, 97, 109, 111, 112, 117, 126, 130, 316, 318, 319, 322, 323, 324, 336, 339, 444, 446, 449, 450, 452, 458, 477, 523, 740, 770
Lasonen-Aarnio, Maria 443, 458
Lave, R. E., Jr. 300, 311
Leblanc, Hugues 251, 263, 276, 684, 702, 703
Lebowitz, J. L. 599
Lederberg, J. 199
Lehmann, Daniel 570, 692, 694, 695, 697, 702
Lehmann, E. L. 300, 310, 335, 339
Lehman, R. 380
Lehrer, Keith 540
Leibniz, Gottfried Wilhelm 36, 59, 61, 65, 67, 317
Leitgeb, Hannes 4, 372, 373, 380, 679, 683, 687, 692, 693, 696, 698, 702
Lele, S. 386, 404
Lennox, James G. 620
Leslie, John 323, 646, 750, 770
Lettieri, A. 274
Levi, Isaac 297, 300, 301, 304, 308, 311, 359, 380, 429, 438
Levin, L. A. 450, 458
Levine, H. P. 692, 699, 300, 308
Lévy, Paul 101, 102, 103, 106, 108, 109, 111
Lewis, C. 404
Lewis, Clarence Irving 777, 786
Lewis, David K. 20, 30, 162, 166, 169, 182, 250, 265, 267, 270, 271, 276, 363, 424–35, 438, 452, 454, 455, 457, 458, 462–70, 474, 630, 631, 632–6, 646, 666, 675, 677, 680, 686, 691, 694, 697, 702, 703, 707, 708. 728, 736, 775–80, 786, 794, 801, 813, 814, 816, 822, 825, 826, 828, 831
Lewis, H.D. 570
Lewis, Peter 678, 680
Lewontin, Richard C. 619
Lexis, Wilhelm 92
Li, M. 214, 233, 246, 440, 444, 450, 459, 530, 534
Liberman, V. 509
Lichtenstein, S. 30, 482, 484, 487, 493, 494, 505, 508, 509, 516, 517
Lieb, E. H. 448, 459
Lin, Hanti 683, 703
Lindeberg, Jarl Waldemar 111
Lindley, Dennis V. 139, 150, 180, 198, 199, 371, 380, 394, 404, 515
Lindsay, R. K. 188, 199
List, Christian 3, 4, 521, 526, 527, 536, 537, 539, 540
Liu, L. 312
Loeb, P. 770
Loève, M. 241, 246
Loewer, Barry 424, 426, 427, 429, 431, 434, 435, 436, 437, 438, 439, 455, 459, 467, 468, 469, 470, 472, 474, 599, 646
Luce, R. Duncan 213, 246, 727, 736, 814
Ludwig, G. 166, 555, 571
Lui, C. H. 516
Łukasiewicz, T. 300, 311, 700
Lundberg, Filip 107
Luzzatto, S. 455, 459
Lyapunov, Aleksandr Mikhailovich 91, 98, 100, 101, 111
Lyon, Aidan 2, 17, 18, 29, 30, 161, 166, 205, 341, 461, 462, 470, 472, 474
Lyons, John 782, 787

Macdonald, R. R. 42, 49
MacFarlane, John G. 252, 255, 276

MacGregor, D. 508, 516
Machina, M. 799, 813
MacKenzie, D. A. 113, 128
Mackey, G. W. 555, 559, 560, 571
Madansky, A. 528, 540
Madow, W. G. 312
Magidor, M. 692, 694, 695, 697, 698, 702
Magnello, M. E. 120, 128
Maher, Patrick 212, 217, 246, 364, 372, 380, 628, 630, 646, 683, 703, 798, 813
Makinson, David 294, 683, 684, 688, 689, 697, 701, 702, 703
Malament, David B. 591, 599
Mann, H. B. 312
Mann, W. E. 648, 771
Manor, O. 244
Manski, C. F. 302, 311
Manson, N. 642, 646
Markov, Andrei Andreevich 94, 95, 98, 101, 103, 105, 108, 111
Marsico, T. 570
Martin-Löf, Per 150, 442, 445, 447, 448, 449, 451, 459
Maryks, R.A. 37, 38, 49
Matthen, Mohan 608, 620
Maty, M. 70, 86
Maudlin, Tim 255, 276
Maxwell, Grover 662, 702
Maxwell, James Clerk 95–7, 111, 177, 573, 574, 576–9, 581, 593, 597, 598, 599
Mayo, Deborah G. 341, 359, 383, 390, 396–401, 404, 650, 661
Mayr, Ernst 606, 620
Mazumdar, M. H. 117, 122, 128
McCarthy, David 4, 710, 722, 723, 724, 727, 736
McClennen, Edward 736, 759, 770, 813
McCloskey, Deirdre N. 393–5, 404, 405
McConway, K. J. 524, 525, 529, 540
McCurdy, Chris 170, 182
McDermott, J. 188, 199
McGee, Vann 174, 182, 692, 703, 784, 787
McGrew, Lydia 646, 743, 747, 748, 749, 770
McGrew, Tim 646, 743, 747, 748, 749, 770
McKay, M. D. 516
McMullin, Ernan 741, 770
Meacham, Christopher J. G. 646, 669, 672, 680, 813
Medina, B. de 37
Meegan, D. V. 483, 484, 494
Megiddo, N. 686, 687, 700
Mehrtens, Herbert 98, 111
Melbourne, I. 459
Mellor, D. Hugh 33, 49, 409, 411, 422, 424, 439, 647
Mendelssohn, Moses 66, 68
Menges, G. 300, 311
Menzies, Peter 828, 831
Méré, Antoine Gombault, Chevalier de 46, 65
Merkhofer, M. W. 499, 513, 517
Merlin, Francesca 601, 620
Meusnier, N. 46, 49, 57, 67
Meyer, M. A. 158, 166, 517
Mikkola, K. 730, 724, 736
Mill, John Stuart 439
Miller, D. W. 422
Mills, Susan K. (also Susan Finsen) 604–12, 614, 615, 616, 618, 619, 620
Millstein, Roberta L. 4, 601, 604, 606, 607, 613, 614, 617, 619, 620, 621
Milne, Peter 265, 276, 656, 661, 787
Miranda, E. 305, 307, 309, 311
Mittelstaedt, P. 547
Mode, E. 209, 246
Molchanov, I. 304, 311
Mongin, Philippe 525, 541, 686, 702, 736
Montague, Richard 455, 459
Montmort, Pierre Rémond de 68, 86
Monton, Bradley 764, 767, 770
Moral, S. 536, 541
Morgan, Charles 684, 703
Morgan, Mary S. 404, 620
Morgan, Millet Granger 503, 513, 515, 517
Morgenstern, Oskar 365, 381, 496, 713, 732, 733, 737, 791, 814
Morris, P. A. 511, 517, 536, 541
Moss, Sarah 627, 647, 672, 675, 677, 678, 680, 783, 787
Mosteller, Frederick 246, 487, 494
Mougin, Gregory 758, 765, 770
Mulaik, S. A. 404
Mullet, E. 489, 494

Mundici, D. 258, 272, 276
Murphy, A. H. 494, 507, 517, 484
Myerson, R. 721, 722, 736
Myrvold, Wayne C. 4, 467, 470, 471, 472, 474, 574, 579, 591, 592, 595, 598, 599
Myung, I. J. 429, 439

Nagel, Thomas 729, 736
Nardini, C. 402, 404
Narens, Louis 225, 246, 727, 736
Nash, R. 73, 86
Nau, R. 306, 311
Neapolitan, Richard E. 2, 185, 187, 188, 192, 193, 195, 196, 197, 199, 820
Nehring, Klaus 536, 541
Neter, E. 507, 515
Neumaier, A. 302, 311
Newell, B. R. 481, 491, 494
Newman, J. R. 487, 488, 494
Newton, Isaac 70, 71, 72, 73, 84, 86, 435
Neyman, Jerzy 119, 124, 125, 126, 128, 148, 347, 382, 386–8, 389, 391, 398, 399, 403, 405
Nicolas of Autrecourt 41
Ng, C. T. 525, 539
Ng, Y. 736
Nies, A. 444, 445
Niiniluoto, I. 328, 338
Nilsson, N. J. 301, 311, 683, 703
Nisticò, G. 570
Niven, W. D. 577, 579, 599
Noordhof, Paul 828, 832
Norton, John 453, 459, 462, 474, 650, 661

Oakes, M. 392, 405
Oakley, J. 494, 515
Oas, G. 555, 571
Ockham, William of 41
O'Hagan, Anthony (also Tony) 383, 405, 485, 492, 494, 507, 515
Ohrenstein, R. 49
Ok, E. 713, 736
Olin, Doris 335, 339
Olivi, P. J. 49
Oppy, Graham 770
Oresme, N. 43, 49
Ortiz, N. R. 514, 517
Osherson, Daniel N. 459, 486, 494, 654, 662
Otsuka, Michael 723, 736
Owen, G. E. L. 339

Paccaut, F. 459
Paccioli, Fra Luca 56, 68
Paley, William 742, 770
Paltiel, O. 244
Papadopoulos, Anthony 614, 621
Parascandola, Mark 826, 832
Parfit, Derek 711, 714, 715, 720, 721, 723, 736
Pargetter, Robert 437, 473
Parikh, Rohit 688, 689, 699, 702
Paris, Jeff B. 134, 249, 261, 272, 276, 703
Parker, Thomas, 1st Earl of Maccesfield 81
Parmentier, M. 65, 68
Pascal, Blaise 2, 45, 46, 47, 49, 53, 56, 57, 65, 66, 68, 316, 738, 770, 789, 790, 812
Pasler-Sauer, J. 516
Pauker, S. 246
Pauly, M. 536, 541
Pearl, Judea 13, 30, 120, 185, 188, 199, 692, 701, 703, 819, 820, 821, 832
Pearle, P. 436, 439
Pearson, Egon Sharpe 113, 124, 125, 128, 382, 386–8, 389, 391, 398, 399, 404, 405
Pearson, Karl 112–4, 116–126, 127, 128, 129, 337, 386, 420
Peirce, Charles Sanders 422
Pepys, Samuel 70, 71, 72, 86
Peters, E. 482, 495
Peterson, D. 680
Pettigrew, Richard 372, 373, 380, 474, 680
Pettit, Philip 540, 813
Phillips, B. 484, 493
Phillips, D. 593, 597
Phillips, L. D. 481, 494, 505, 509, 517
Piccione, M. 680
Pidgeon, N. F. 486, 492
Piggins, A. 537, 540
Pigliucci, Massimo 608, 621

Piron, S. 44, 49, 57, 67
Pitcairne, Archibald 70, 72, 73, 74, 75, 86
Pitowsky, Itamar 564, 565, 571
Pitt, M. A. 439
Plancherel, Michel 98
Plantinga, Alvin 742, 770
Plato 317, 339
Poincaré, Henri 98, 101, 148, 294, 320, 321, 339, 599
Poisson, Siméon-Denis 50, 77, 79, 91, 94, 95, 101, 315, 339, 347, 470, 471, 591, 599
Pólya, Georg 108, 338
Popper, Karl Raimund 134, 149, 150, 160, 163, 164, 166, 172–8, 180, 182, 251, 179, 180, 295, 389, 397, 405, 406–9, 411–6, 418, 419, 422, 429, 439, 461, 474, 574, 600, 612, 621, 650, 661, 684, 688, 689, 703
Por, H.-H. 244, 491, 495
Porter, C. 451, 457, 459
Porter, T. M. 113, 129
Postkasse 246
Potter, J. M. 301, 311
Prade, H. 274, 700
Pratt, J. 246, 275
Price, Huw 438, 782, 787
Price, Richard 70, 82, 83, 86, 323, 615, 738, 770
Prierias, Sylvester 37, 49
Priest, Graham 20, 30, 276, 769
Provine, W. B. 113, 129
Ptolemy 39, 49, 316
Puppe, C. 536, 541
Pust, Joel 679, 680
Putnam, Hilary 273, 565, 571, 649

Quaeghebeur, E. 309
Quiggin, John 309
Quine, Willard Van Orman 635, 647, 656, 782, 787, 801

Rabinovitch, N. L. 49
Rabinowicz, Wlodek 723, 736
Radon, Johann 100, 178
Raftery, A. 385, 404
Raiffa, Howard 484, 490, 509, 515, 529, 541
Rakow, T. 481, 494, 515
Ramsey, Frank Plumpton 3, 131, 133, 134, 138–44, 148, 149, 150, 168, 182, 299, 319, 323, 339, 340, 362–6, 370, 380, 424, 439, 477, 478, 494, 627, 628, 647, 681, 703, 707, 708, 713, 717, 736, 763, 770, 774, 777, 787, 799–801, 813
Ramsey, Grant 616, 618, 621
Ramsey, Joseph 830, 831
Randall, C. H. 555, 571, 572
Ranjan, R. 512, 517
Rapoport, Anatol 489, 494, 496
Rathmanner, S. 440, 459
Ratto, M. 518
Raven, J. E. 317, 339
Ravinder, H. V. 503, 517
Rawls, John 711, 716–9, 722, 724, 725, 734, 736, 737
Reagan, R. T. 211, 246, 487, 494
Real, L. A. 480, 494
Redhead, Michael 462, 474
Regoli, G. 224, 246
Reichenbach, Hans 130, 133, 134–8, 144, 145, 148, 149, 150, 151, 341, 342, 343–6, 348, 349, 352–6, 358, 359, 815, 816, 820, 832
Reichenbach, Maria 151
Rényi, Alfred 160, 166, 172, 173, 175–8, 180, 182
Rescher, Nicholas 340, 661, 688, 703, 704, 770, 813, 831
Resnik, Michael 713, 718, 736
Reznikova, Z. 207, 247
Rice, Sean 614, 615, 621
Richardson, M. 683, 703
Richardson, Robert C. 604, 606, 607, 621
Rimini, A. 438, 593
Rios, S. 301, 306, 307, 310
Rissanen, J. 192, 199, 428, 429, 439
Ritov, I. 490, 494
Robartes, Francis 53, 70, 72, 74, 75, 76, 77, 79, 84
Roberts, F. 76, 86, 727, 736
Roberts, John T. 427, 438, 463, 474, 634, 647
Robins, James 826, 831
Robinson, Abraham 162, 166
Robinson, J. A. 701

Rode, C. 490, 494
Roeper, Peter 251, 163, 276, 684, 703
Romeijn, Jan Willem 388, 392, 405
Rosenberg, Alex 606, 619, 621
Rosenblatt, J. 30, 307, 308
Rosenhouse, J. 334, 339
Rosenthal, Arthur 98
Rosenthal, J. 321, 339, 592, 600
Roseveare, N. 374, 380
Ross, H. 741, 770
Ross, Jacob 678, 680, 683, 703
Rothman, Kenneth J. 342, 359, 824, 826, 832
Rothschild, Daniel 778, 787
Rottenstreich, Y. 482, 483, 492
Royall, Richard 382, 386, 389, 401, 405, 650, 661, 741, 752, 770
Royden, H. L. 770
Rubinstein, A. 677, 680
Ruggeri, F. 300, 302, 304, 310
Runde, J. 411, 422
Russell, Bertrand 203, 250, 269, 772
Russell, Gillian 702
Russell, Paul 739, 770
Russo, Federica 621, 831
Russo, J. E. 484, 494
Ryabko, B. 207, 247

Sadrolhefazi, A. 247, 303, 311
Saffiotti, A. 700
Sagan, Carl 217, 247
Sagrado, J. 536, 541
Sahlin, Nils-Eric 301, 306, 309, 812
Salmon, Merrilee H. 149
Salmon, Wesley C. 136–8, 145, 149, 151, 158, 166, 348, 353, 354, 356, 359, 416, 421, 422, 652, 653, 661, 662
Saltelli, A. 500, 518
Sanfilippo, G. 700
Santerna, P. 44, 49
Santorio, Paolo 669, 675, 680
Sargent, Thomas J. 301, 310
Satia, J. K. 300, 311
Saurin, Joseph 53
Savage, C. Wade 379
Savage, Leonard J. 143, 144, 150, 277, 295, 301, 309, 365–7, 369, 381, 400, 401, 403, 405, 471, 474, 490, 507, 518, 574, 592, 600, 628, 645, 647, 707, 708, 717, 724, 736, 791, 792, 794, 801, 809, 812, 813
Scanlon, Thomas 712, 719, 729, 737
Schack, R. 244
Schaffer, Jonathan 429, 431, 439, 455, 459, 465–7, 469, 470, 474, 828, 832
Scheffler, Samuel 714, 715, 737
Scheines, Richard 199, 819, 821, 823, 830, 831, 832
Schervish, Mark J. 166, 178, 182, 311, 379, 381, 401, 404, 489, 494, 529, 540, 678, 680, 770
Schick, Frederic 364, 369, 381, 629, 647
Schlesinger, George 759, 770
Schmeidler, D. 301, 306, 307, 310, 311, 813
Schmidt, U. 726, 737, 806, 814
Schmidt-Petri, C. 275, 379, 380, 438, 645, 646, 683, 702
Schneider, Ivo 65, 68
Schnorr, C. P. 450, 451, 459
Schoemaker, P. J. 484, 494
Schoenberger, A. 487, 491
Schrenk, M. A. 437, 439
Schroeder, Mark 683, 703
Schulz, Moritz 669, 680
Schurz, Gerhard 294, 692, 693, 703
Schüssler, R. 37, 49
Schwartz, D. 37, 49, 245, 457
Schwarz, Wolfgang 4, 361, 433, 434, 439, 440, 463, 474, 633, 647, 669, 680, 707, 737
Schweitzer, M. E. 489, 494
Scott, Dana 227, 247, 259, 276, 683, 703
Scozzafava, R. 301, 309
Scriven, Michael 605, 621
Searles, D. 244
Segale, C. 488, 495
Segerberg, Krister 689, 704
Seidenberg, A. 246, 275
Seidenfeld, Teddy 160, 166, 180, 182, 306, 308, 311, 379, 381, 401, 404, 489, 494, 629, 647, 678, 680, 770, 798, 814
Sellke, T. 393, 403
Sen, Amartya 711, 718, 725, 737
Seneta, E. 113, 128
Sennet, Adam 787

Sextus Empiricus 40, 41
Shachter, R. D. 188, 199, 503, 518
Shackle, G. L. S. 299, 311
Shafer, Glenn 62, 68, 100, 111, 160, 166, 205, 228, 229, 245, 247, 258, 276, 298, 300, 302, 304, 311, 312, 479, 495, 643, 648
Shafir, S. 490, 496
Shakespeare, William 44–5
Shalizi, C. 402, 403
Shannon, Claude E. 587, 600
Shapiro, Stewart 450, 459
Sheps, Mindel 826, 832
Shimony, Abner 363, 381, 563, 567, 570, 572, 629, 630, 647
Shoesmith, E. 318, 339
Shortliffe, E. H. 188, 199
Siegler, M. V. 393, 404
Simmonds, R. 689, 703
Simon, Herbert A. 478, 479, 495
Simpson, Edward 11–4, 30
Simpson, R. H. 487, 495
Simpson, Thomas 70, 82, 86
Skipper, Robert A. 604, 621
Sklar, Lawrence 359, 591, 600
Skyrms, Brian 20, 30, 182, 329, 338, 339, 340, 363, 364, 375, 380, 381, 502, 518, 628, 630, 647, 704, 733, 737, 766, 771, 786, 794, 814, 815, 818, 832
Slinko, A. 244
Slovic, Paul 30, 367, 380, 478, 482, 485, 490, 492, 493, 495, 507, 508, 515, 516, 517, 518
Small, M. 515, 517
Smets, Philippe 228, 700
Smets, Sonia 689, 700
Smith, A. F. M. 384, 403, 516
Smith, C. A. B. 299, 300, 305, 311, 479, 495
Smith, D. 812
Smith, E. E. 486, 494
Smith, Michael 734, 735
Smith, Nicholas J. J. 248, 250, 256, 276
Smith, Peter 455, 459, 700
Smithson, Michael 3, 4, 165, 479, 483, 488, 489, 490, 491, 495
Smolin, Lee 742, 771
Snir, M. 428, 438, 447, 458, 683, 701
Snow, P. 692, 694, 704
Soames, Scott 250, 276
Sobel, Howard 762, 771
Sober, Elliott 606, 609, 610, 612, 613, 614, 616, 617, 621, 622, 641, 648, 741, 742, 743, 745, 746, 749–54, 758, 759, 765, 767, 770, 771
Soll, J. B. 484, 493, 495, 517
Solomonoff, R. 440, 459
Sopena, A. 483, 495
Sorbo, L. 584, 597
Sorensen, Roy 760, 771
Spearman, C. 129
Specker, E. P. 555, 563–5, 567, 569, 571, 572
Spector, B. 776, 786
Speranski, S. O. 699, 704
Spetzler, C. S. 500, 503, 513, 518
Spiekermann, Kai 536, 540
Spielman, S. 158, 166, 389, 390, 391, 405
Spirtes, Peter 193, 197, 199, 819, 821, 830, 831, 832
Spohn, Wolfgang 536, 541, 694, 704, 792, 814
Sprenger, Jan 3, 4, 341, 392, 395, 401, 402, 404, 405, 740, 771
Stael von Holstein, C-A. S. 500, 502, 503, 509, 513, 518
Staffel, Julia 379, 381
Stalker, D. 326, 340
Stalnaker, Robert C. 168, 182, 250, 270, 276, 626, 630, 635, 648, 675–7, 680, 694, 697, 704, 777, 778, 786, 787, 793, 812, 814
Stanhope, Philip, 2nd Earl of 82
Stanovich, K. E. 478, 495
Steiger, J. H. 404
Stieltjes, Thomas Jan 94, 99, 100
Stigler, Stephen M. 61, 63, 68, 71, 72, 73, 75, 80, 83, 86, 112, 113, 114, 118, 123, 129, 322, 333, 340
Stirling, James 80, 82
Stone, D. R. 488, 495
Stone, M. 511, 518, 523, 541
Strevens, Michael 321, 340, 434, 439, 470, 472, 474, 592, 600
Strode, Thomas 70, 73, 78, 84, 86

Student (William Sealy Gosset) 120–4, 129, 386
Sturgeon, Scott 640, 648, 683, 704
Suárez, Francisco 37, 38, 49
Suárez, Mauricio 660
Sudderth, William D. 336, 339
Sugden, R. 726, 737, 806, 814
Suppes, Patrick 146–8, 149, 150, 151, 245, 246, 300, 312, 422, 438, 474, 555, 571, 600, 662, 684, 693, 699, 703, 704, 727, 736, 815, 832
Swain, M. 646, 683, 704
Swanson, Eric. 4, 782, 783, 784, 787
Swijtink, Zeno G. 318, 340
Swinburne, Richard 739, 742, 750, 761, 771
Swirles, Bertha 150
Sylla, Edith D. 2, 45, 49, 51, 54, 56, 62, 63, 64, 65, 66, 67, 68, 73, 74, 78, 85, 337
Sylvester, Prieras 37. 49

Tarantola, S. 518
Taylor, H. A. 485, 496
Teigen, K. H. 244, 491, 495
Teller, Paul 265, 363, 381, 634, 648
Temkin, Larry 726, 737
Tenenbaum, Joshua B. 215, 247
Teng, Choh Man 311, 637, 646, 656, 661, 830, 831
Tentori, Katya 3, 4, 175, 652, 654, 656, 658, 659, 660, 662, 740, 769
Terwijn, S. A. 688, 704
Thomas, T. 722, 730, 736
Thomson, James F. 775, 787
Thomson, Judith Jarvis 709, 711, 733, 737
Thomson, T. 83, 86
Thomson, William 577, 600
Thrush, Michael J. 642, 646
Thuring, M. 486, 495
Tillers, P. 209, 247
Tipler, F. J. 641, 644, 741, 768
Titelbaum, Michael G. 4, 635, 648, 667, 669, 672–5, 677, 678, 680
Todd, P. M. 493
Todhunter, Isaac 76, 86
Tolman, R. C. 589, 596, 600
Tooby, J. 487, 491, 494
Tooley, Michael 435, 439
Toti Rigatelli, L. 46, 49
Toulmin, Stephen E. 36, 48, 782, 787
Tranel, D. 302, 310
Trauth, K. M. 501, 518
Tribe, E. L. 47, 49, 209, 247
Troffaes, M. C. M. 298, 302, 305, 307, 308, 309, 312
Tsao, C. J. 488, 495
Tukey, John W. 300, 312
Turing, Alan 327, 445, 449, 451, 654
Tversky, Amos 246, 367, 380, 478–82, 485–7, 490, 492, 493, 495, 496, 507–10, 515, 516, 517, 518, 727, 736, 807, 813
Twardy, Charles 828, 832

Uffink, Jos 467, 474, 575, 578, 589, 598, 600
Urbach, Peter 158, 166, 335, 338, 364, 375, 380, 384, 404, 410, 411, 422, 637, 645, 740, 769

Vadhan, S. P. 443, 459
Vailati, Giovanni 142
Valla, L. 36, 40
Vallentyne, Peter 792, 814
van de Ven, A. H. 510, 516
van der Waerden, B. L. 53, 67, 68
van Fraassen, Bas C. 7, 30, 160, 166, 174, 182, 251, 267, 276, 333, 340, 352, 359, 363, 364, 371, 376, 381, 426, 434, 439, 629, 631, 632, 640, 648, 678, 680, 684, 686, 704, 740, 771
van Hees, M. 536, 541
van Lambalgen, M. 448, 449, 450, 459
Varadarajan, V. S. 555, 572
Varignon, Pierre 54
Venn, John 132, 204, 213, 238, 247, 323, 340, 341, 343–6, 348, 349, 358, 359
Verkuilen, J. 495
Verma, Thomas 820, 832
Vicig, P. 302, 312
Vickrey, W. 715, 737
Ville, Jean 106, 447, 448, 455, 459
Vineberg, Susan 678, 680
Vitanyi, Paul M. B. 214, 233, 246, 440, 444, 450, 459

von Fintel, Kai 780, 787
von Kries, Johannes 92, 298, 309, 319, 321, 323, 338, 340, 591, 600
von Mises, Richard 99, 100, 101, 108, 111, 132–5, 138, 150, 151, 164, 197, 199, 341, 343–9, 352, 354–8, 359, 406, 407, 409, 419, 420, 422, 438, 442, 444–9, 451, 452, 455, 457, 459
von Neumann, John (also 'Johann') 203, 246, 365, 381, 478, 496, 555, 570, 593, 713, 732, 733, 737, 791, 814
von Plato, Jan 88, 98, 111, 321, 340, 470, 474, 576, 591, 600
von Smoluchowski, Marian 108
von Winterfeldt, D. 513, 514, 515, 517
Voorhoeve, Alex 723, 736
Vovk, Vladimir 100, 111, 205, 247, 302, 311
Vranas, Peter B. M. 464, 474, 634, 648

Wagner, Carl G. 377, 378, 381, 524–6, 539, 540, 541
Wakker, Peter 726, 737, 814
Wald, Abraham 111, 126, 129, 446, 459
Waldegrave, Francis 77, 78, 84
Wallace, Alfred Russell 602
Wallace, David 536, 572, 602, 635, 645
Walley, Peter 163, 166, 211, 216, 218, 222, 224, 231, 232, 234, 243, 247, 297, 301, 305–8, 312, 479, 488, 496, 640, 648
Wallner, A. 312
Wallsten, T. S. 224, 487, 488, 489, 491, 494, 496
Walsh, Denis M. 613, 622
Walsh, Sean 457
Wang, Ucilla 208, 247
Waters, Kenneth C. 606, 622
Wattenberg, F. 428, 437, 631, 644
Weatherson, Brian 255, 256, 258, 276, 626, 643, 648, 677, 680, 785, 786, 787
Weber, E. U. 480, 488, 493, 496
Weber, Marcel 149, 473, 603, 619
Weber, Martin 479, 491
Weber, Tullio 438, 593
Wedgwood, Ralph 683, 704
Weichselberger, K. 303, 306, 312
Weinstein, S. 459
Weisberg, Jonathan 643, 648, 743, 750, 771, 784, 787, 808, 813
Welch, E. 45, 49
Werndl, Charlotte 454, 459
West, R. F. 478, 495
Westbrooke, I. 12, 30
Weymark, J. 536, 541, 718, 737
Wheeler, Greg 638, 648, 688, 704
Wheeler, T. A. 517
White, Roger 640–3, 648, 811, 814
Whitworth,W. A. 340
Wilce, Ales 555, 570, 572
Wiener, Norbert 98, 100, 109
Williams, B. 731, 736, 737
Williams, J. Robert G. 2, 20, 249, 261, 262, 265, 276
Williams, Mary B. 606, 621
Williams, P. M. 301, 305, 312
Williamson, Jon 139, 149, 151, 158, 166, 363, 381, 384, 404, 405, 621, 635, 638, 648, 649
Williamson, Timothy 3, 162, 166, 248, 249, 250, 271, 276, 458, 631, 637–9, 648
Wilson, Alastair 166, 439
Wilson, Bradley E. 613, 620
Wiman, Anders 99
Winkler, R. L. 484, 494, 501, 506, 509, 512, 515, 517, 518
Winman, A. 485, 493, 496, 508, 518
Winsberg, Eric 437, 439
Wittgenstein, Ludwig 132, 235, 236, 247, 338, 647, 772
Wolfenson, M. 235, 247, 301, 312
Wolff, R. W. 512, 515
Wolford, G. 485, 496
Wolpert, R. L. 384, 402, 403
Woodward, James 822, 832
Wooff, D. A. 302, 310
Wright, Crispin 250, 255, 276, 643, 644, 647, 649
Wright, Sewall 123, 819, 832
Wu, G. 480, 493

Yager, R. 300, 312
Yalcin, Seth 211, 218, 247, 689, 704, 774, 782, 783, 787, 788
Yaniv, I. 484, 485, 496

Yates, J. F. 300, 312, 486, 496
Younes, Z. 245
Young, M. L. 516
Youtz, C. 246, 487, 494
Yudkowsky, Eliezer 700
Yule, G. U. 11, 30, 112, 120, 121, 125, 126, 129
Yushkevich, A. P. 54, 58

Zabell, Sandy L. 3, 16, 131, 151, 317, 319, 323, 327, 329, 332, 333, 337, 340, 363, 379, 452, 591, 599, 640, 649
Zadeh, Lofti A. 188, 216, 256, 265, 276, 479, 496
Zaffalon, M. 244, 302, 309
Zanotti, Mario 147, 151
Zellner, Arnold 139, 151, 182
Zhang, Jiji 830, 831
Zidek, J. V. 511, 516, 521, 529, 540
Ziliak, Stephen T. 393–5, 404, 405
Zukowski, L. G. 300, 312
Zwick, R. 496
Zynda, Lyle 3, 299, 312, 369, 374, 376, 378, 379, 381, 384, 405, 628, 629, 649, 740, 771, 795, 799, 814

Subject Index

ability
  relation to manifestation 604–6, 612, 616
  reproductive 603–6, 609, 612, 617, 618
  survival 603–6, 616, 618
abstract formulae 20
accuracy 39, 60, 118, 138, 183, 204, 212–5, 217, 232, 243, 292, 355, 371–3, 442, 484–5, 504
  arguments 261–3, 271–3, 678
  Brier Score explication of 143–4, 261 (n. 24), 371, 489
  domination 261–3, 629–30
  and precision 204, 212–7, 232, 234, 243, 293, 484
achievement 716, 718, 728
action-guiding 710, 796 (n. 13)
Adams' thesis 168, 682, 684, 689–93, 774, 776 (n. 4), 777 (n. 7), 779–80
additivity
  additivity axiom 6, 284, 351
  axiom of dual additivity 257
  countable, sigma, σ- 18, 105, 156, 205
  finite 6, 156, 159–60, 205, 351
  subadditivity 234, 259, 510
aggregation
  attitude 537
  judgement 536–7
  opinion 510, 512, 519–21, 525, 526, 536, 539
  utility 537, 706, 721, 728
algebra (see also 'field') 17
  algebraic analysis 57
  Boolean 555–60
  sigma (σ-) 17, 21, 100, 105, 156, 163–4, 167, 176, 205–6, 217, 239, 460–1
  Heyting 270
aleatory contracts 43–5
Allais' paradox 367, 810
ambiguity aversion 300–2, 303, 308, 349, 483, 490
annuities
  life 44
anonymity
  and expert judgement 510
  and utility aggregation 720
anti-realism 143–5, 248, 270, 427, 431
assertibility 775–776, 777 (n. 7), 780
astronomy, evidence in 39, 42
attitude
  de dicto 676
  de re 675–6
  de se (see also 'credence, self-locating') 635–6, 666
  doxastic 625
  propositional 250
averaging, axiom of 269
axiom of choice 18, 20, 157
axioms 6, 17
axiomatization of probability
  Kolmogorov 17–19, 156, 351, 356
  nonclassical 256–7, 259
  Popperian 160, 173, 251 (n. 7), 279–82
axiomatizations, immanent versus logical 272

backgammon 74
base rate 9–10, 390
  neglect 10, 11 (n. 4), 390, 392–3, 507
Bayes factor 377, 378, 385
Bayes' theorem 8–9, 15, 18, 24, 169, 183–5, 266, 286, 287, 384, 511, 738, 745, 752, 768
Bayesian method 88, 91, 131, 136, 139, 143, 207, 341 (n. 2)
Bayesian network (Bayes net)
  causal 189–91, 815, 819–20
  definition 184–5

Bayesianism 382–6
objective 136, 139, 145, 335, 384, 440, 592 (n. 17), 635
types of 132, 139, 142, 145, 146, 147, 386, 402, 440, 635
behaviour, as manifestation of a propensity 133–4, 408–9, 606, 611, 612, 617, 619
belief
conditional 779, 780
degree of (see also 'credence') 33, 133, 138, 142, 143, 161, 168, 180, 212, 229, 253, 255, 261, 267, 322, 330, 360–2, 365, 371–3, 379, 382, 384, 393, 394, 423, 461, 477, 502, 779, 802
full (binary) 155, 253, 255, 626, 627, 767
partial 155, 255, 262, 625–7
rational constraints on 627–37
revision (see also 'conditionalization', 'probability kinematics', 'updating') 333
belief function
Dempster-Shafer 160, 228, 258, 300, 304, 643
Berkson's paradox 13–4, 16
Bernoulli trials 75
Bernoulli's hypothesis 790
Bernoulli's theorem (see also 'law of large numbers') 50
best system analysis 4, 423–37
bet (betting) 45–6
one-side 305
betterness (for a person, individual, relation) 706, 711, 714–30
bias (biases)
anchoring 481–3, 486, 503, 508
catch-all underestimation bias 482
cognitive 478–9, 481, 483, 485, 489, 500, 501, 503, 507–9, 513–4
confirmation 481
in judgment formation 507–10
motivational 501, 513
overconfidence/apparent overconfidence 484–5, 509
packing 508
prior distribution 746, 754, 755–7
process 10, 169, 172, 385, 389, 407, 408, 446–7, 452, 462, 595, 712, 751 (n. 24)
pruning 508
sampling (selection effect) 14, 190, 191, 395, 400, 481, 485
sequence 444–6, 448
zero sum 483–4
Bienaymé-Chebyshev inequality 90, 93, 94, 103, 104, 105
billiards 83, 593
bills of mortality, London, 39, 75
Biometrika 119, 121–5
biometry 117–126
binomial distribution 23, 28, 29, 71, 321, 419
normal approximation to 77, 79, 82, 83
Poisson approximation to 77, 79
binomial expansion 62, 78, 79, 80
binomial method of the unciæ 76
binomial theorem 70, 72, 75–6
biological practice 618
birth rates 75, 88, 318
bivalence 249
blame 708
Boltzmann equation 106, 573–4
Boltzmann's H-function 576, 578, 581–3
borderline cases 250
Borel paradox 179
Borel set 21
bounded rationality 479
Brier score 143–4, 261(n. 24), 371, 489
Brownian motion 98, 100, 107–9

calibration 226, 371, 372 (n. 15), 484–5, 499, 505–7, 511, 629–30
chart 499, 505
graph 505
cardinal utility 365, 713, 791
casuistry 36–8, 45
causal discovery 822, 823 (n. 4)
causal graph 189, 820, 822
causal inference 819, 822
causal model 819–22
causal power 826, 826 (n. 7)
causal relevance 817–8
causal strength 815, 823, 824, 825
causality (see also 'causation') 41, 133, 137, 141, 145, 148, 415–6, 568
deterministic 415–6, 427, 820
indeterministic 826
quasi-deterministic 576–8

causation (see also 'causality') 91, 116, 416, 423, 430, 434, 435, 793, 815–30
  actual 827–30
  probabilistic 4, 91, 148, 188, 417, 815–30
cause, interacting 817–8
centered proposition 635–6, 669–70, 671–7
centered world 666–8, 675
central limit theorem (see also 'limit theorems') 29, 87, 100–1
certainty 255
  anti-certainty 255–6
  degree of (see also 'credence', 'degree of belief') 36, 51, 60, 318
  moral 36–7
certainty factor 188
certainty-loss framework 674
chance (see also 'interpretations of probability, best system', 'interpretations of probability, objective', 'probability, objective') 4, 169–70, 423–437
  reductionism 440, 442, 452, 456–7
  and determinism 453–5
Chapman-Kolmogorov equation 107
characteristic function 22
chi-squared (χ^2) 119, 123, 386
Choquet capacities 303
Church of England 72
circularity problem 605–6, 608, 610, 618
circulation of blood 70, 72
classical mechanics, classical physics 163, 453–5, 462–3, 467, 545, 575, 580, 593, 652
  indeterminism in 453–4
  chaotic dynamics in classical statistical mechanics 454
classical theory of probability 3, 77, 452
cognitive loading 255, 261, 262, 263, 272, 273
coherence
  and accuracy 371–3
  probabilistic 363–4, 368–9, 429, 628–30, 800 (n. 22)
  and rationality 373–4
  strict (see also 'regularity') 363, 631
coin tossing 147, 420
collective, *Kollektiv* 99
comparative probability 224–31, 277–9, 299
completeness
  axiom 225, 300
  informational 147
  interpersonal 716–9, 725, 726, 729
  logical 257–60, 681, 685–7, 689, 692, 697–8
  rule 289, 291
compositionality 773, 775, 777, 783, 785
concentration 102
conditional 168–9
  counterfactual 710, 774, 794
  indicative 682, 691, 693, 699, 774–6, 778, 779–80
conditional assertion 779–81
conditional excluded middle 777
conditional independence 11, 185–6, 195, 821
conditional probability 7–9, 18, 34, 88, 100, 156–7, 160–2, 167–81, 184–6, 188, 206, 263–7, 272–3, 279, 281–2, 323, 327, 356, 361, 363, 384, 392–3, 397, 461, 485–7, 631, 637–9, 684, 686, 688–9, 691–2, 694–5, 774–9, 784, 793, 815–6
  comparative 278–82, 284, 293
  fallacies involving 485–7
  and non-classical logics 263–7, 272–3
  in quantum mechanics 550, 553
  and conditionals 168–9, 181, 682, 684, 774–9, 784
conditionality principle 385
conditionalization, conditioning (see also 'updating') 9, 11, 146, 170, 174–7, 204, 214, 248, 263–5, 279–85, 303, 306–7, 361–4, 374–7, 384, 402, 404, 431, 503, 633–8, 667–78, 750, 753, 755
  compartmentalized 672–5
  for imprecise probabilities 306
  for indeterminate probabilities 306
  Jeffrey 316, 332, 361–4, 373, 377, 378, 667, 678
  shifted 670–1

confidence 3, 8, 243, 251–2
  comparative 277–8
  interval 124, 213, 396, 484–5
  over- 484–5, 509
confirmation 10, 175
  absolute 651
  bias 481
  degree of 131, 282, 293–4, 650, 741
  and the fine tuning argument 742–3, 746, 755, 757
  incremental 651
  relations 490
  theory 3, 175, 267, 273, 402, 650–65
conflict aversion 479, 490
conflicting information 479
conglomerability 177–81
conjecture 51, 59–60, 62, 63
conjunction 167, 173, 249, 252, 254, 257, 521, 526
  fallacy 478, 485–7, 490
  probability of 7, 259, 688
connective 257, 259, 264–5, 684, 686, 690–1, 775, 777, 780
consequence
  logical 681–9, 691, 692, 695–7, 699
  no-drop characterisation of 257–8, 274
  nonclassical logical 252–6, 273
  rational 283
  special consequence condition 652
consequentialism 706, 711–2, 714, 797
  act 730–1
  rule 731–3
consistency theorem 100, 109
constraint-based learning 192, 196
constraint semantics 783–5
constraint
  synchronic 627, 631, 632, 635, 636, 666
  diachronic 627, 634, 637, 667
context dependency 144, 146, 387, 486, 488, 603 (n. 3), 778
continuity
  axiom 306, 706, 713, 724, 726, 729
  monotone 205
  semantics 692, 698
contractualism 711–2, 731, 724, 729
contradiction 20, 174, 175, 248, 252, 254, 280–1, 289, 521, 778
convergence
  almost sure 28, 238, 241, 347, 575, 765
  in distribution 25, 95, 99, 100–102, 108, 199
  in probability 28, 88, 105, 139, 204, 214, 318, 347–8, 371, 409, 575, 693
  of relative frequencies 135–138, 143, 199, 204, 233, 235, 241–3, 347, 353, 575
convex combination of truth values 251, 253, 256, 257, 258, 261, 262, 272, 274
correlation 10–6, 41, 104, 116–22, 170, 190–1, 374, 478, 526, 534, 536, 562–4, 617, 658, 793, 816, 818–820, 823, 828
  coefficient, measure 104, 116–22, 564
  conditional 11–4, 16
  positive, negative 10–11, 14, 16, 563, 658, 819
  spurious 216–7, 816, 818, 823, 828
cost-benefit analysis 20, 66
counterfactual (see also 'conditional') 710, 774, 794
  and causation 423, 430, 434–5, 822, 825, 828
  conditional probability 175
  priors 374–5
  probability 470
  probability logic of 693–4
credal sets 304–305
credence (see also 'degree of belief') 8
  base 264
  conditional 173, 595, 627, 631–4, 637, 639, 667 (n. 4), 668, 676, 740
  function 174, 252, 432, 433, 464, 628–32, 633
  fuzzy vs. sharp 640–1, 643
  imprecise 232 (n. 28), 251 (n. 9), 640
  indeterminate 163
  profile 255 (n. 3)
  self-locating 636, 666–78
  unconditional 173, 627, 630, 634, 667 (n. 4), 676 (n. 26), 794
cumulative distribution function 24
  joint 25
cyclical preferences 367–9

de Moivre theorem 83, 107
decision-making with imprecise probabilities 307–8

decision-making with indeterminate probabilities 307–8
decision rule, Jeffrey 268
decision theory 26, 126, 146, 214, 268–9, 273, 301, 366, 370, 384, 387–8, 706–7, 709, 711 (n. 3), 712–3, 730–3, 738, 757–67, 789–814
  causal 175, 732, 794
  evidential (see also 'decision theory, Jeffrey, Jeffrey-Bolker') 732
  Jeffrey, Jeffrey-Bolker (see also 'decision theory, evidential') 268–9, 273, 366
'definitely' operator 260 (n. 23)
degree of belief (see also 'credence') 33, 133, 138, 142, 143, 161, 168, 180, 212, 229, 253, 255, 261, 267, 322, 330, 360–2, 365, 371–3, 379, 382, 384, 393, 394, 423, 461, 477, 502, 779, 802
Dempster rule 300
Dempster-Shafer theory 300
density function 23
deontology 712, 733–4
dependence
  causal 10, 732
  counterfactual 825
  probabilistic 10, 425, 732, 821
  "radical" 778
  sample space (partition) 392, 400, 482–3, 488
determinism 89, 97, 147, 415, 429, 455, 460–3, 465–6, 469, 471–2, 574–5
desirability 268–9, 272
desire 797
  consistency of 799, 800, 802
  interpreting 801–810
  satisfaction 366, 711, 773, 795, 797, 801
dialetheism/dialetheic/dialethic 252
Diamond's example 721–2
dice
  biased, loaded (see also 'bias, process') 10, 74, 407–8
  games 45–6
  dicing problem, probability of the sum of the faces that show 70, 79
diffusion process 107–8
directed acyclic graph (DAG) 184–5, 819
discounting 191
disjunction 17
dispersion 102–3
disposition 4, 408–9, 411, 413, 420, 423, 424, 605–6
  partial 424, 431
distribution
  binomial 23, 28, 29, 71, 321, 419
  definition 22
  identical 22
  joint 22–3, 187, 545, 546, 586
  normal (Gaussian) 24, 29, 89, 90, 94, 95, 101, 115, 118, 120, 121, 330, 419
  of propensities for reproductive success 607, 609–11, 613–6, 618
  uniform 24, 27, 179 (n. 4), 229, 389, 468, 469, 586, 744, 746, 747, 754, 763
distribution function 99, 101–2, 105, 106–107, 392, 499
  cumulative 238–9, 245, 399, 503, 505, 511, 514
  stable 102
division of stakes problem 46, 70, 76, 77
dogmatism 642–3, 650 (n. 2)
dominance 306, 722, 798, 818
  strict 722
  weak 722
Duhem-Quine thesis 656
duration of play problem 77–8
Dutch book
  definition 215
  argument 262–3, 265, 272, 363–4, 369, 379, 628, 634, 678
  theorem 282

e-admissibility 308
effect size 393–4, 401–2
elicitation, probability 210, 214–5, 297, 484–5, 487–9, 497–518, 800–1
Ellsberg's paradox 299
embedding 780, 782, 784
empirical support (see also 'confirmation', 'evidence') 421
empiricism 40, 135, 137, 142, 146, 147
English Statistical School 112–26
ensemble approach to statistical mechanics 95–7, 586–7, 588–9
environment 603–18, 742

epidemiology 342, 416, 497, 815, 823–4
epistemic modal 773
epistemic utility 635
equality (moral) 721–3
  principle 722
equiprobability 47, 75, 88, 92, 99, 315, 317, 356 (n. 18), 456, 556, 566, 569, 588, 717–8, 776
ergodicity 96–8, 108, 448 (n. 4), 583
error statistics 399
error
  theory of 89–90, 93–4, 114, 116–8, 121–3, 125
  type I and type II 387–8, 395, 400
estimation
  of probability or frequency 147, 149, 188–9, 198–9, 209–10, 213, 301, 410, 415, 434, 478, 483–4, 486, 488–90, 502, 507–8, 512, 519
  theory 90, 93–5, 116, 123–4, 330, 391, 396
  of truth 372–3, 773
ethical theory 72, 705
ethics 705–34
eugenics 114, 117, 120, 122, 124, 604 (n. 5)
evaluation
  of an act 707–10
  of an agent 708
event 6, 17, 20
  decision theory 791
  measure theory 17
  single 28, 356, 407, 409–15, 482, 490, 827
  types 20
evidence 10
  admissible 464
  centered 674
  confirming 652
  empirical 565, 584
  historical 40
  imprecise or incomplete 811
  in astronomy 39, 42
  law of 34–6
  legal standard of 825
  measure of 385
  medical 40
  new (see also 'conditionalisation') 634
  problem of old 175, 373–5
  same 171
  statistical 383, 386, 393, 397, 401–2
  total 636–7
exchangeability 131, 142, 143, 302, 306, 307, 316, 324–6, 328
  partial 329
expectation 25–26, 52–8, 61–6, 71–3, 76, 105–6, 178, 365, 369, 506, 576, 587, 593–4, 759–61, 764–6, 790–1, 795–7
  lower and upper 232, 243, 298, 300, 302, 304–5, 307–8
  of truth value 251, 253, 256, 267
expected goodness 714–5
expected utility
  axioms 712–3
  function 365–6, 628–9, 713, 791–2, 795–800
  theory 628–9, 712–3, 791–7, 801
experimental arrangement 134, 408, 412, 548, 567, 569
expert
  judgment, opinion 484, 497–501, 510, 519
  principle 267, 433
  system 188, 300
explanation 137–8, 807
expressivism 691, 780

fairness 55, 706, 721–2, 729
  of a bet/betting system 143, 628, 641
faithfulness condition 192–7, 821, 823
  embedded 197
fallacy
  base rate 10, 11 (n. 4), 390, 392–3, 507
  conjunction 478, 485–7, 490
  gambler's 27
falsehood, logical 251, 254, 257, 262
falsifiability
  methodological 419–21
favouring 11
field (see also 'algebra') 17
  Borel (see also σ-field) 351
  σ- (sigma; see also 'σ-algebra') 17, 18, 24, 27, 164

fitness 601–19
fortune 57, 74, 316–7
fraction
  attributable 825
  etiological 826
  excess 824–6
  population-attributable 826–7
freedom
  degrees of 97, 547–8, 575, 577, 580, 581, 585, 586, 587
  moral/political 711, 725
frequency
  finite 342–3
  hypothetical (limiting relative) 28, 343–58, 444, 448, 453, 575
  relative 28, 43, 64–5, 197–9, 213
frequentism 132–8, 341–58, 382–402, 444, 451–3, 457
  conditional 402
  finite 342–3
  frequentist inference 382, 395–6, 402
  hypothetical 343–58
fuzzy credence 640–1
fuzzy logic 265, 301, 479
fuzzy set, fuzzy membership 216 (n. 20), 258, 486, 488

gambler's fallacy 27
gambling manuals 74, 80
game
  against nature 732
  against other people 732–3
  coordination 731–2
  sequential 732
  simultaneous 732
game of chance
  coin toss, fair 454
  dice 2, 21–3, 33, 45–6, 59, 65
  equity or fairness in (*aequa ratione*) 56
  premature ends of, division of stakes 46, 58, 70, 76–7, 790
game theory 478
generating condition 134, 408
generating function 77, 79, 89
genetics 113, 121, 124, 467, 468, 601, 604, 607
God 2, 42, 738–40
  God's foreknowledge 61, 66
  pragmatic arguments for belief 757–67
  probabalistic arguments for 738
goodness of fit test 119, 386
gradational accuracy 371–2

half-proof 35–6
Harsanyi's theorem 718–21, 725, 727, 728–9
hazard
  environmental 823–5
  jeux de 53
heredity 113–6, 118–23
heuristic (see also 'bias')
  affect 482
  anchoring-and-adjustment 481–2
  availability 482, 507
  frugal 479
  representativeness 507
higher-order probability (see also 'second-order probability') 685–6
historical evidence 40
Humphreys' paradox 161, 169–71, 415–7, 417 (n. 7)
hypothesis of elementary errors 90
hypothesis test (see also 'significance test') 42, 387–8, 391, 402
hypothetico-deductivism 402

ideal agent 249, 270, 274, 282, 707, 730, 802, 809
identification, causal 823
impartiality 717–9, 721–2
implicature 776, 781
imprecise probabilities 171–3, 181, 297–312
IncExc axiom 257, 258, 259, 260, 265 (n. 30)
incommensurability 724
incompleteness 712, 717
indices of guilt, in Roman law 35
independence
  axiom 306, 366
  causal 732, 821
  conditional 11, 185–6, 195, 821
  definition 10
  eventwise 524–7
  generalisation 19
  for imprecise probabilities 303, 307

independence (*Cont.*)
and identically distributed (i.i.d.) trials or variables 23, 27, 29, 91, 93, 233, 237–8, 240, 242, 329–30, 347, 352, 385, 389, 394, 451, 456
for indeterminate probabilities 303, 307
of infinite collections of events 10
probabilistic 10, 11, 19, 22, 143, 267, 302, 512
of random variables 22, 105
strong 307, 712–3, 721–2
indeterminacy 155, 163, 203–4, 210, 214, 216, 232–4, 235, 297–308
indeterminate probabilities 297–312
indeterminism 453–5
indexicals 670–1, 676
indicator function 22
indifference principle 3, 319–20, 322, 588
induction
by enumeration 135–7, 323, 325, 331, 353
Goodman's new riddle of 656
principle of 135
problem of 40–2
rule of 353
inductive logic 131–2, 657, 683
inference
causal 821–2
inductive 41, 47, 323, 325, 386, 401–2
logical 249, 697, 699
probabilistic 183, 387
statistical 42, 87, 114, 122, 143, 328, 341 (n. 2), 382–402, 420–1, 428
infinitely divisible distribution 103, 108
information cost 731, 799
informativeness 10, 425, 427–8, 436, 484, 505–6, 715–6, 746, 752, 801, 804–5
inheritance, biological 113–6, 118–23
insurance
pricing of 43–5, 53
interpretations of probability 3–4
best system 4, 423–437
classical theory 3, 77, 452
frequency/frequentist 3–4, 132, 134–5, 341–58, 382–402, 442, 575
finite 162–3, 342–3
hypothetical 343–58, 444, 448, 453, 575
logical 3, 131, 360, 374, 384
propensity theory 406–21
subjective (see also 'credence', 'probability, types of, subjective') 360–379
objective (see also 'chance', 'probability, types of, objective') 92, 146–7, 320, 456, 461, 576–7, 626, 708
intersection 19, 164, 167, 205 (n. 2), 521
intuitionism 20, 251 (n. 7), 258
irrationality 726
Islamic law
of contracts 43
of inheritance 56
of partnerships 56

Jeffrey conditionalization 316, 332, 361–4, 373, 377, 378, 667, 678
Jesus of Nazareth 73
just noticeable difference 715

kinetic theory of gases 95–8, 573, 578–9
knowledge 51, 133–7, 140–8, 188, 206 (n. 6), 318–9, 323, 349 (n. 11), 498–500, 574–5, 594–6, 625, 627, 637, 687
Kochen-Specker theorems 270, 555, 561, 563, 564–5, 569
Kolmogorov axioms 6, 17–8, 100, 107, 155–65, 174, 177–81, 205–6, 269, 351, 356, 417, 420–1, 549–51, 627–8
Kolmogorov complexity 456
prefix-free complexity 450
compressibility 449
Kolmogorov Zero-One Law 236, 240

label space 99, 521, 639
Laplacian approximation method 88
law
court trials 47, 91, 208–9, 825 (n. 5)
Islamic 43, 56
legal arguments 34–6, 47
of torts 825
Roman 35–6, 43
law of excluded middle 250–2, 254, 273
law of iterated logarithm 104
law of large numbers 26–7, 50–1, 60–3, 66
empirical 344, 346–7

strong 26, 238
weak 27
law of total probability 7, 18
lawn bowling 77
leases on lives 81
least squares 89–90, 93
Lebesgue measure 447
Leviathan 72
leximin 717
Lindley's paradox 394–5, 396, 399, 400 (n. 13), 401
life expectancy 53, 410
life table 81–2
likelihood 8
law of 655, 741, 744
likelihood principle 384, 384–5, 641–642
maximum likelihood method 90, 122–3
ratio 386, 388, 399, 654
likelihoodism 386, 650 (n. 1), 738, 752–4, 767
limit theorems 27–9, 87–89, 91, 95–5, 98, 100–1, 106
Lindeberg condition 102
logic 20, 248–76, 681–704
algebraic 270–2
choice of 255
classical 20, 155, 157, 168, 248–51, 257, 259 (n. 19), 270, 273, 565, 683–5
default 699
fuzzy 265, 301, 479
intuitionist 20, 251 (n. 7), 258
Łukasiewicz 258
many-valued 265, 536–7
modal 218–24, 688
paraconsistent 20, 252 (n. 10), 253
probabilistic 300–2, 681–99
quantum 248 (n. 1), 270, 273, 565
revision of 565
strong Kleene 254
supervaluational 249 (n. 2), 251, 252, 258, 259 (n. 19), 260, 265 (n. 30), 266, 269, 725
symmetric 254, 258
three-valued 253–6
logicism 130–2
long run 27–8, 83, 213–4, 238, 241–43, 382, 387, 396, 412–4, 418–21, 611, 790
lot (*sors*) 54–7
lower expectations 304–5
lower previsions 305
lower probabilities 303–4

Markov chain 105–6
Markov condition 186–7, 189–98, 819–21
causal 820–1
Markov property 108
Martingale 106
maximality 308
maximal options 308
maximin 717
measurable sets 17
measure theory 16–19
measurement
evaluative 726–8
theory 204, 215
medical evidence 40
memory loss (see also 'credence, self-locating', 'sleeping beauty problem') 678 (n. 32)
mentalese (mental language) 250, 270
method of moments 117–22
minimality condition 821
minimax options 308
minimum description length (MDL) principle 192
miracles 40, 41, 83, 738
mismatch problem 605–6, 610, 618
modal
epistemic 773
logic 249, 258, 688
probability 218, 223–4, 423
metaphysical 463, 470
modus ponens 776
moment problem 94, 99
moments 611, 614
method of 117–22
theorem 322
theory of 94–5
money pump 369
monotonicity, axiom of 251, 252, 256, 699
cautious 697
Monty Hall paradox 14–6
Myerson's example 721–2

natural selection 123, 374, 601, 603–4, 606 (n. 9), 607, 610, 613–4, 616, 617, 618
nearest and dearest objection 731
necessity 28
  contractual 63
  hypothetical 63
  institutional 63
  physical 63, 423
negation 17
negation rule 6
neutral existence 720
Newcomb's problem 732, 793, 794
Newtonian mechanics 163, 453–5, 462–3, 467, 545, 575, 580, 593, 652
non-expected utility theory 726
nonidentity problem 720
non-measurable sets 18, 20
non-monotonicity 283, 651–2
non-negativity, axiom of 6, 156, 163, 205, 251
normal distribution 24, 29, 89, 90, 94, 95, 101, 115, 118, 120, 121, 330, 419
normalization, axiom of 156, 158, 233, 251, 269
normalization objection 747–9
normativity 711
normed sums 102–3, 105–6
"not by chance" arguments 42
null hypothesis (see also 'hypothesis test', 'significance test') 387–95, 397, 401, 448, 637

objective chance, objective probability (see also 'chance', 'probability, objective' 'interpretations of probability, objective') 249, 271, 328, 360, 441, 455, 463 (n. 7), 463–5, 468, 574–5, 632–4, 639
objectivity 134, 137, 140, 143, 357–8, 387, 401–2, 431
occupancy problem 77
odds 499, 502
offspring 115, 118, 605, 606, 607–10, 614–5, 617, 618
old evidence, problem of 175
omniscience 61, 253, 256, 373, 375, 739
omniscience, credal 256
one-sided betting 305
one-sided testing problem 391–3, 399
ordering
  complete 306, 307, 712
  comparative probabilistic 224–5
  partial 306, 308, 556
  preorder 712, 725
  probabilistic 694
  problem of 453
organism 601–19
ought 709–10, 730, 790–1, 795–800
outcome 20
  fecund and sterile 62
  space (see also 'sample space') 21, 22, 28, 329, 453, 546, 740
overconfidence 484–5, 509

p-values 383, 391–402
paraconsistency 20, 252 (n. 10), 253
paradox 11–16
  Allais' 367, 810
  Berkson's 13–14, 16
  Borel 179
  Ellsberg's 299, 307, 811
  for finite additivity 159–60
  Humphreys' 161, 169–71, 415–7, 417 (n. 7)
  Lindley's 394–5, 396
  lottery 688
  Monty Hall 14–16
  preface 626
  purported paradox in probability 333–6
  raven 656
  Robartes's (Roberts's) 75–6
  Simpson's 11–13
  sorites 123, 212, 232, 248 (n. 1)
  St. Petersburg 26–7, 66, 790
parameter estimation 391
Pareto 718
  ex ante 708–9
  equality-neutral 722
parish clerks, company of 75
partial entailment 657–9
partiality 717–9, 721–2
partition 7
  infinite 18
  partition dependence 482–3, 488
  principle of 260 (n. 22, 23)

partnerships or companies 55–7
penalty methods 144 (n. 28)
permutation postulate 131, 324, 326, 329
permutations and combinations 50, 52, 53, 58, 73, 231, 453, 718
phase space 96, 467, 580–3, 585–9, 593, 595
place selection 132, 344–6, 445–6, 452, 477
pluralism 147, 402, 468, 657, 659
pool, problem of the 77–8
population
  size 613–4, 706
  statistics 88
  variable 719–21
possibility measures 304
power set 17, 241, 521, 690
pragmatic loading 262, 263, 273
pragmatism 141–6, 431
  non-pragmatism 370
preemption 826–8, 830
preference 794–796
  partially ordered 306
  rational 628–9, 707, 712–3, 726, 789, 799
  satisfaction 716, 729, 730
preorder 712, 725
presumption 35–6
prevention 818, 824
primary intensions 20
principal principle 432–5, 463–5, 632–4, 639
principle of indifference 319–20, 322, 588, 740, 764
prior probability 8–9, 139, 140, 184–5, 293, 383, 384, 387, 528, 530, 534, 538–9, 740, 755–6
  improper 176–77
  reference prior 402
prioritarianism (priority view, priority to the worse-off) 705, 723–4, 727, 729
probabilis 51
probabilism 261–3, 369, 373–4, 627–30, 640–1
  moral theological 37–8
  radical 142–8
probabilistic logic 300–2, 681–99
probability
  axioms of 6–11
  bearers of 19–20, 269–71
  convex-combination characterisation of 253
  decomposition of assessment of 503
  measure of 205, 213, 223–4, 226, 231, 304, 546, 550–1, 555–6, 585, 682–9, 747, 756
probability, interpretations of 3–4
  best system 4, 423–437
  classical theory 3, 77, 452
  frequency/frequentist 3–4, 132, 134–5, 341–58, 382–402, 442, 575
    finite 162–3, 342–3
    hypothetical 343–58, 444, 448, 453, 575
  logical 3, 131, 360, 374, 384
  propensity theory 406–21
  subjective (see also 'credence', 'probability, types of, subjective') 360–379
  objective (see also 'chance', 'probability, types of, objective') 92, 146–7, 320, 456, 461, 576–7, 626, 708
probability of specific occurrences
  causation 825–6
  death 44, 342, 354, 407, 507–8, 818, 825 (n. 5)
  ruin 78
probability, types of
  additive (see also 'additivity') 6, 284, 351
  comparative 224–231, 277–9, 299
  conditional 7–9, 18, 34, 88, 100, 156–7, 160–2, 167–81, 184–6, 188, 206, 263–7, 272–3, 279, 281–2, 323, 327, 356, 361, 363, 384, 392–3, 397, 461, 485–7, 631, 637–9, 684, 686, 688–9, 691–2, 694–5, 774–9, 784, 793, 815–6
  epistemic (see also 'belief, degrees of', 'credence', 'interpretations of probability, subjective', 'probability, types of, evidential') 33–4, 319
  evidential 3, 271, 440, 637–40, 782
  extrinsic and intrinsic 38
  frequentist (see also 'frequency', 'frequentism', 'interpretations of probability, frequency/frequentist') 132–8, 341–58, 382–402, 444, 451–3, 457
  higher-order (see also 'second-order') 685–6

probability, types of (*Cont.*)
imprecise 297–312
indeterminate 297–312
inverse 118, 122, 125, 139
logical (see also 'interpretations of probability, logical') 3, 131, 360, 374, 384
lower 303
nonclassical 256–61
objective (see also 'chance', 'interpretations of probability, best system', 'interpretations of probability, objective') 92, 146–7, 320, 349, 356, 410–1, 456, 461, 574–5, 626, 708
posterior 8–9, 384
prior 8–9, 139, 140, 184–5, 293, 383–4, 387, 528, 530, 534, 538–9, 740, 755–6
quantum 545–69
qualitative 2, 35 207, 277
second-order 502, 685
single-case (singular) 4, 343–4, 354, 407, 412–5, 451–3, 574
stochastic/factual 33, 51, 93, 101, 103, 105, 106–9, 126, 237, 428, 441, 548
sub-additive 234, 259, 369, 510
subjective (see also 'credence', 'interpretations of probability, subjective') 134, 143–4, 199, 210, 226, 256, 263, 274, 320, 360–79, 410, 477–8, 479, 482–4, 484–5, 486, 503, 707 (n. 1), 707–8, 759, 765–7, 775, 791, 795, 797, 800
that sum to more than one 162
unquantified 47
upper 303
verbal 487–9
probability elicitation 210, 214–5, 297, 484–5, 487–9, 497–518, 800–1
probability function 17, 19, 28, 138, 162, 167–81, 277, 293, 316, 538–9, 592, 631, 640, 650, 707–8, 791–2
probability intervals 147, 211–2, 302–3, 488
probability judgement 497–514
probability kinematics (see also 'conditionalization', 'updating') 332–3, 361
probability space 17, 21, 156, 375–6, 378, 545, 560, 561, 563, 569, 773, 777, 816
infinite 17, 100
probability theory, generalizations of 555–60
probability tree 503
probability weighting function 479–80
probability wheel 499, 502, 513, 514
propensity theory of probability 406–21
proportional syllogism (statistical syllogism or argument from frequencies) 42–3
propositions 6, 20
fine-grained vs. coarse-grained 250, 269–71
prospect theory 367, 806, 807
cumulative 480

quality of experience 711, 716
quality of life 711
quandary 255 (n. 13)
quantum mechanics 545–572
Copenhagen interpretation 677 (n. 29)
Everettian interpretation 677 (n. 29)

radical interpretation 728
raffle 74, 76, 772–3, 782–3
Ramsey test 774, 777
random drift 601, 606, 618
random variable
definition 21, 178 (n. 3)
range of 21
unbounded 26
vector-valued 21
randomization (see also 'randomness, process') 316
randomized control experiment (RCE) 189
randomness
algorithmic 444
as evidence for chance 456–7
as typicality 446–9
as ungoverned by a rule 444–6
process 442
product 441
pseudorandomness 443
random sequences (see also 'randomness, product') 344–5, 442–3

random walk 448
reduction of chance to 442, 452, 457
rank-dependent expected utility theory 480
rational preferences 628–9, 707, 712–3, 726, 789, 799
rationality (see also 'irrationality') 143, 224–5, 250–3, 260–1, 361, 366, 373, 634–5, 726, 796–803, 807–11
bounded 479
Humean 707–8, 797, 805
real number values 18, 21, 162–3, 165, 213–4, 252–3, 255, 263, 366, 740
extended 760, 762–3
recarving, principle of 259–60, 272
recurring series 78
reference class problem 42, 354–6, 638
reflection principle 631, 678, 686
regularity (see also 'coherence, strict') 251, 363 (n. 8), 630–1, 650 (n. 2), 662, 667, 740, 740 (n. 4), 763
axiom of 162, 363
regression 116, 118–20, 122
rejection
and credence 255–6
and significance 387, 389, 400, 419
relevance-limiting thesis 673–4
reparation rule 376–8
repeatable conditions 409, 412–5, 418, 611
representation theorem 143, 145, 273, 291, 324, 328–30, 365–6, 369, 791–5
reproductive success 605–6, 608–15, 618
repugnant conclusion 721
resources 725
responsibility 708, 827
reversions (see also 'regression') 81–2, 166
risk
attitudes to 66, 628, 790
difference 824–6
and ethics 711, 714–6, 720, 723, 724, 727–9
evaluations of 482, 490, 497, 823
and insurance 43–5
ratio 824–6
relative 824
versus uncertainty 298–9, 503, 717
robust Bayesian statistics 300
robustness/resiliency 297–300, 776
robustness analysis 300
Royal Oak lottery 74
Royal Society 76, 82–3, 114, 119
Royal Statistical Society (also Statistical Society of London) 113–4, 120–1
ruin, probability of 78
rule of mixture 52
runs, problem of 79

Saint Petersburg paradox 26–7, 66, 790
sample space (see also 'outcome space', 'dependence, sample space') 105, 205, 690
satisficing 479
screening off 655 (n. 8), 816, 818–20, 829
second-order probability 502, 685
selection bias 14, 190–1, 395, 400, 401, 485
self-locating credence 636, 666–78
self-reference 248, 733
semantics
classical 249, 279 (n. 5)
constraint 783–5
nonclassical 249–50, 253–6
for high probability conditionals 690–6
truth conditional 777–9
sentences 19
set theory 20
sets of states 20
severity 388, 396–9
sex ratio 75
share (*pars*, *portio*) in division of stakes problem 79
sharpenings 725
σ-algebra (sigma algebra) 17, 21, 100, 105, 156, 163–4, 167, 176, 205–6, 217, 239, 460–1
σ-field (sigma field) 17
significance test (see also 'hypothesis test') 115, 118, 388–94, 396, 401, 448
null hypothesis significance test (NHST) 387, 389, 391, 394, 397
simplicity (of theories) 39
Simpson's paradox 11–13
single case
problem of 4, 343, 412, 414
skew-symmetric bilinear (SSB) theory 367–9

sleeping beauty problem 636, 677–8
social contract 711
social welfare
  ex ante 708–9
  ex post 708–9
SRI protocol 513–4
Stalnaker's thesis 777–8
standard deviation 26–7
statements 20
statistical evidence 383, 386, 393, 397, 401–2
statistical inference 87, 382–3, 394, 397, 402, 428–9
statistical physics 87, 92, 97, 98
Statistical Society of London (also Royal Statistical Society) 113–4, 120–1
statistical syllogism (proportional syllogism or argument from frequencies) 42–3
stochastic convergence (see also 'law of large numbers') 105–6
stochastic process (see also 'randomness, process') 47, 100, 103, 106–9, 126, 237
  with continuous time parameter 106
  with independent increments 107
stochastics 51
Stieltjes integral 99, 100
stopping
  optional 383, 385 (n. 4), 399
  rule 400–1
subjectivism (see also 'credence', 'interpretations of probability, subjective', 'probability, types of, subjective') 3, 133, 360–2
sufficiency principle 385
supertruth/supertrue 252 (n. 9), 258
supervaluation (see also 'logic, supervaluational', 'sharpenings') 172, 173, 181, 249 (n. 2), 251, 252, 258, 259(n. 19), 260, 265(n. 30), 266, 269, 725
supposition 774, 777, 781, 793, 794
suspension of judgement 255, 317
Szpilrajn's theorem 725

Talmud 35, 43
tautology (see also 'truth, logical') 254
tennis
  handicaps in, calculation of 63
  probability applied to 63
time average 97, 204, 236, 237, 238, 241, 591
topics, in rhetoric, as ways or organizing arguments 59
torture 35
total evidence 168, 170, 273, 333, 464, 633 (n. 28), 638–9, 640
  requirement of 636–7
transmission, transmission failure 643–4
triviality results 169, 691, 777, 779
trolley problem 733, 734
truncation of random variable 101
truth
  conditions 272, 775, 777, 779, 782
  deflationism concerning 255
  logical (see also 'tautology') 251, 361, 373
truth-functional connective 775, 776
truth status
  cognitive loading of 255, 261, 262, 263, 272, 273
  Kleene truth status assignment 253, 254, 255, 256, 257, 258, 264, 265, 267, 274
  pragmatic loading of 262, 263, 273
truth values
  convex combinations of 251, 253, 256, 257, 258, 261, 262, 272, 274
  designated 256, 258
  drop in 256, 272
  estimation of 372–3, 773
  expectation of 257
truth-value gaps 249 (n. 2), 250
typicality
  measure theoretic approach to 446–448, 452
  effective measure one properties and 447

unanimity 519, 523, 524, 525, 528, 529, 530, 534, 537, 709
uncertainty (see also 'credence', 'interpretations of probability, subjective') 297–8
  assumption 717–8
union 17, 164, 205 (n. 2), 351, 521
uniqueness thesis 639
unpredictability 443, 454

updating (see also 'conditionalization') 9, 144, 204, 263, 361, 363, 364, 364 (n. 10), 528, 530 (n. 9), 531, 538, 635, 667–76, 677–8, 765, 766
  on de se information, on centered propositions 635–6, 669–77
    demonstrative scheme 669, 675–7
    shifting scheme 669–71
    stable base scheme 671–5
upper expectations 304–5
upper previsions 297, 305
upper probabilities 303
utilitarianism 706, 718, 726, 727, 731, 732, 733
utility
  epistemic 635
  expected 365–6, 628–9, 712–3, 791–7, 801
  marginal 365
  non-expected 726
universal set 6

vagueness 172, 271, 293, 297, 361 (n. 1), 488–9, 626, 640 (n. 44), 689, 725
validity
  classical 249
  nonclassical 248–9
variable
  dependent 108, 116
  independent 101
  observed 193
  random (see also 'random variable') 21–2
variance 26
veil of ignorance 716–9, 722, 724, 729
Venn diagram 7, 9

Waldegrave's Problem 77–8
weighing 479–81
weight 51, 131, 136
welfare 711
welfare economics 708, 728
Whist 74
Wiener measure 109
Wiener space 109
would-be 408–9

zero axiom 251, 252, 256, 257, 259
zero-fit problem 428
Zero-One Law 236, 240
zero-set property 511