

Conditionals, Paradox, and Probability

Conditionals, Paradox, and Probability

*Themes from the Philosophy
of Dorothy Edgington*

EDITED BY

Lee Walters and
John Hawthorne

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

© the several contributors 2021

The moral rights of the authors have been asserted

First Edition published in 2021

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: 2019946131

ISBN 978-0-19-871273-2

Printed and bound in the UK by
TJ Books Limited

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

<i>List of Tables</i>	vii
<i>List of Contributors</i>	ix
1. Introduction <i>Lee Walters</i>	1
2. Philosophy and Me <i>Dorothy Edgington</i>	8
3. A Note on Conditionals and Restrictors <i>Daniel Rothschild</i>	19
4. Chasing Hook: Quantified Indicative Conditionals <i>Angelika Kratzer</i>	40
5. New Paradigm Psychology of Conditional Reasoning and Its Philosophical Sources <i>David Over</i>	58
6. Counterfactuals to the Rescue <i>Cleo Condoravdi</i>	76
7. Counterfactuals and Probability <i>Robert Stalnaker</i>	100
8. Grammar Matters <i>Sabine Iatridou</i>	122
9. Constructing the Impossible <i>Kit Fine</i>	141
10. The Epistemic Use of ‘Ought’ <i>John Hawthorne</i>	164
11. Undercutting Defeat and Edgington’s Burglar <i>Scott Sturgeon</i>	174
12. Edgington on Possible Knowledge of Unknown Truth <i>Timothy Williamson</i>	195
13. Prefaces, Sorites, and Guides to Reasoning <i>Rosanna Keefe</i>	212
14. Hysteresis Hypotheses <i>Alan Hájek</i>	227
15. Verities and Truth-values <i>Nicholas K. Jones</i>	239
<i>Bibliography of works by Dorothy Edgington</i>	263
<i>Index of Names</i>	267
<i>Index of Topics</i>	268

List of Tables

5.1. The ‘defective’, 2×2 de Finetti table for <i>if p then q</i>	62
5.2. The 3×3 general de Finetti table for <i>if p then q</i>	65
5.3. The Jeffrey table for <i>if p then q</i>	66

List of Contributors

CLEO CONDORAVDI is Professor of Linguistics at Stanford University.

DOROTHY EDGINGTON is Waynflete Professor of Metaphysical Philosophy Emeritus at the University of Oxford.

KIT FINE is University Professor and Silver Professor of Philosophy and Mathematics at New York University, and Distinguished Research Professor at the University of Birmingham.

ALAN HÁJEK is Professor of Philosophy at the Australian National University.

JOHN HAWTHORNE is Professor of Philosophy at the University of Southern California.

SABINE IATRIDOU is Professor of Linguistics at the Massachusetts Institute of Technology.

NICHOLAS K. JONES is Official Fellow of St John's College and Associate Professor of Philosophy at the University of Oxford.

ROSANNA KEEFE is Professor of Philosophy at the University of Sheffield.

ANGELIKA KRATZER is Professor Emerita of Linguistics at the University of Massachusetts Amherst.

DAVID OVER is Emeritus Professor of Psychology at Durham University.

DANIEL ROTHSCHILD is Professor of Philosophy at University College London.

ROBERT STALNAKER is Professor Emeritus of Philosophy at the Massachusetts Institute of Technology.

SCOTT STURGEON is Professor of Philosophy at the University of Birmingham.

LEE WALTERS is Associate Professor of Philosophy at the University of Southampton.

TIMOTHY WILLIAMSON is Wykeham Professor of Logic at the University of Oxford.

1

Introduction

Lee Walters

Dorothy Edgington's work has been at the centre of a range of ongoing debates in philosophical logic, philosophy of mind and language, metaphysics, and epistemology. This work has focused, although by no means exclusively, on the overlapping areas of conditionals, probability, and paradox. In what follows, I briefly sketch some themes from these three areas relevant to Dorothy's work, highlighting how some of Dorothy's work and some of the contributions of this volume fit in to these debates.

1.1 Conditionals

Often we face deep-rooted uncertainty, and so the best we can do is to estimate the probabilities involved, rather than making outright judgments as to the truth or falsity of a claim. For example, there are ten balls in a bag, five red and five white; Priya picks an unseen ball from the bag at random; has Priya picked a red ball? The prudent answer is not to affirm or deny outright that Priya has picked a red ball, but rather to say that it is 50% likely that she has picked a red ball (and consequently, 50% likely that she has not).

This retreat to probabilistic judgements from outright affirmations and denials is not limited to categorical claims such as Priya picked a red ball. It is also present in our consideration of conditional statements. So, adding to the previous example, let's say that three of the red balls have black spots. What should our attitude be to the claim that if Priya picked a red ball, it had a black spot? Well, as three of the five red balls have black spots, the appropriate answer seems to be that it is 60% likely that if Priya picked a red ball, it had a black spot.

Considerations of this sort make attractive the claim that the probability of a conditional is equal to the conditional probability of its consequent on its antecedent.

* Dorothy Edgington's work far outstrips the three topics discussed above. And her contributions to the three debates mentioned are more numerous and offer more insight than can be discussed here. Nevertheless, from this brief overview, and from the chapters contained in the rest of the volume, we can see the ingenuity and the breadth of Edgington's work, the difficulty of the problems that she has focused on, and how she has advanced our understanding of them.

The Equation: $p(\text{if } A, C) = p(C/A)$,

where $p(C/A) = p(A \& C)/p(A)$, when $p(A) \neq 0$.

So far, so good. The problem is that David Lewis (1976) proved that there could be no proposition expressed by ‘if A , C ’ that satisfied The Equation, and so conditionals were not evaluable as true or false, if The Equation holds. So shocking was this result, given the intuitiveness of The Equation, that Robert Stalnaker described Lewis’s result as a ‘bombshell’ in a letter to Bas van Fraassen. This is because Stalnaker was trying to give an account of the propositions expressed by conditional statements that respected The Equation. In the face of this dilemma, Stalnaker rejected The Equation and maintained that conditionals express propositions (see, *inter alia*, his contribution to this volume). Others, like Edgington, held on to The Equation and denied that conditionals expressed propositions.

The Equation is extremely intuitive, so in the face of Lewis’s proof, why not give up the assumption that conditionals express propositions as Edgington does? First, we can ask what are people doing when they accept and put forward conditionals, if not believing and asserting them? The non-propositionalist can respond that we conditionally believe and conditionally assert conditionals, where these notions may not be reducible to further mental states or speech acts. From a propositionalist perspective this response may seem *ad hoc*, but, as Edgington points out, there are not only what, from her perspective, are conditional assertions, but also conditional commands (if it rains, take in the sun loungers) and conditional questions (if Liverpool score first, will they win?), where these are not obviously equivalent to outright imperatives and questions.

A second objection to non-propositionalism states that whereas one might be able to accept that indicative conditionals (if Priya took a red ball, it has a black spot) do not express propositions, it is too much to accept that counterfactual conditionals (if Priya had taken a red ball, it would have had a black spot) do not express conditionals. But whereas Stalnaker reasons from the claim that counterfactuals express propositions to the claim that indicatives also express propositions, Edgington reasons in the opposite direction providing analogous reasons for endorsing non-propositionalism about counterfactuals as she does for indicatives.

Perhaps the main reason for accepting that conditionals express propositions is that the alternative appears to be subject to a version of the Frege–Geach problem that plagues expressivism.¹ That is, whatever non-propositional meaning we assign to conditionals has to be consistent with the meaning of conditionals when they are embedded in more complex linguistic forms. We can, for example, not only embed categorical statements, statements that straightforwardly express propositions, in the consequents of conditionals, we can also seemingly embed conditionals in the consequents of other conditionals, resulting in structures such as ‘if A , then if B , then C ’. But if what is required to be the consequent of a conditional statement is a proposition when the consequent is a categorical claim, presumably a proposition is required when the consequent is itself a conditional. Similar points can be made with other operators such as conjunction, disjunction, negation, and modal operators.

¹ See Williamson’s chapter in this volume for discussion.

A simple way of accounting for the embedding of conditionals in more complex linguistic structures is in terms of propositional contents. But if we reject that conditionals express propositions, we need to find some other way of accounting for embedding conditionals. As it stands, this is a challenge to those who want to hold on to The Equation, rather than a proof that The Equation has to be rejected. Still, the onus appears to be on those who maintain The Equation, like Edgington, to provide a compositional account of the meanings of conditionals in non-propositional terms that allows them to be embedded.

The above construal of the dialectic rests on the assumption that we can unproblematically embed conditionals in more complex linguistic forms. But this is far from obvious. For example, (i) 'if *A*, then if *B*, *C*' seems equivalent to (ii) 'if (*A* & *B*), *C*'. But if (i) and (ii) are equivalent, then as Gibbard (1981) has shown, 'if *A*, *C*' is equivalent to $\sim A \vee C$ and so the falsity of a conditional's antecedent is sufficient for its truth. But this consequence unacceptable. To take one of Edgington's examples, from the fact that The Queen is not at home, it does not follow that if The Queen is at home, she is waiting for me to telephone!²

Moreover, as McGee (1985) argued, treating embedded conditionals as expressing propositions seems inconsistent with modus ponens. For example,

1. If a Republican wins, then if Reagan does not win, Anderson will.
2. A Republican will win.

Therefore,

3. If Reagan does not win, Anderson will.

(3) seemingly follows from (1) and (2) by modus ponens. But whereas (1) and (2) both seem true, (3) does not (the Democrat Carter was running second in the polls, with Anderson a distant third). McGee takes such cases to show that modus ponens is indeed invalid, but such a conclusion is hard to swallow. The non-propositionalist can note that she does not run into this difficulty, for on her construal, the argument above should be replaced with:

4. If a Republican wins and it is not Reagan, then Anderson will win.
5. A Republican other than Reagan wins.
6. Anderson will win.

And this argument is valid. The propositionalist, then, needs to explain why we are inclined to accept that (i) and (ii) are equivalent when, for him, they are not.

More generally, Edgington has argued that embeddings of conditionals are problematic, a result that is surprising, if they express propositions. Angelika Kratzer (Chapter 4, this volume) furthers the debate on embeddings of conditionals by considering what account we can give of quantified conditionals such as 'no one will pass, if they goof off'. Kratzer argues that such conditionals do represent a

² Gibbard takes his proof to show that indicative conditionals do not express propositions. In the same paper, he also presents his example of Sly Pete, a so-called 'Gibbard case' to reason to the same conclusion. Edgington (1997) argues, however, that the proper lesson of such cases is that indicative conditionals are somehow epistemic, rather than that they do not express propositions.

problem for a propositional account of conditionals and that there is no general account of the embeddings of conditionals. This is music to Edgington's ears, but rather than rejecting propositionalism, Kratzer appeals to pragmatics to address the problem (see also Chapter 3, this volume, where Rothschild argues that by adopting a Kratzer-style treatment of 'if' as a restrictor, propositionalists have the resources to respond to a number of arguments for non-propositionalism).

1.2 The Paradox of Vagueness

A second literature to which Edgington has made important contributions and which is the focus of several chapters in this volume, is the paradox of vagueness. Many concepts expressed in natural language appear to be vague, in the sense that they appear to lack precise application conditions and admit of borderline cases, cases where it neither seems right to say the concept applies nor to say that the concept does not apply. For example, consider the concept *bald*. Although there are people who are clearly bald and people who are clearly not bald, there are people who are in-between, neither clearly bald nor clearly not bald. It will not do, apparently, to say that such people are neither bald nor not bald, since, plausibly, this is to contradict oneself. The trouble does not end there, however.

It is characteristic of vague concepts, unlike precise concepts, that they appear to obey a principle of tolerance, the iterated application of which leads to absurdities such as that someone with a million hairs on their head is bald.³ The paradoxical reasoning only requires two premises and multiple applications of modus ponens:

7. A person with zero hairs on their head is bald.
8. If a person with $n-1$ hairs on their head is bald, a person with n hairs on their head is bald.

Therefore,

9. A person with one hair on their head is bald.

(7) is an obvious application of the concept *bald*; (8) is justified by the principle of tolerance that appears to be characteristic of vague concepts; and (9) follows from (7) and (8), by modus ponens. By itself, (9) does not represent a problem. But once we have (9), we can combine this with (8) to derive that a person with two hairs on their head is bald, and so, by repeated applications of (8), we can deduce that someone with a million hairs on their head is bald!

How should we respond to such paradoxical reasoning? Edgington (1996) approaches the paradox of vagueness by arguing that there is analogy between it and the preface paradox. In particular, she thinks that we can learn lessons about vagueness by considering the degree of belief response to the preface paradox.

The preface paradox is as follows. A careful author believes each of the claims she makes in her book, but, acknowledging her fallibility, she states in the preface that some

³ In our toy example, I'm ignoring the fact that whether someone is bald depends not only upon how many hairs they have on their head, but also upon the distribution of those hairs.

claims she makes in the book are bound to be false. The author appears to be rational, but we can reason from a number of claims that she takes to be true (each claim she makes in the book) to a claim that she rejects, namely that all the claims in the book are true. What to do? The degree-theoretic response is to say that the author believes each individual claim she makes in the book to some high degree less than certainty, but that when she considers the book as a whole her individual doubts add up, so that she believes the entire book only with a low degree of certainty. This approach can be formalized by modelling degrees of belief probabilistically to show that the probability of the conclusion of a valid argument can be lower than the probability of any of the premises. Certainty, unlike truth, is not preserved by valid reasoning.

Edgington's account of vagueness, like the above approach to the preface paradox, employs a probabilistic degree-theoretic structure that she calls 'verities' or 'degrees of closeness to clear truth'. How does this help with the paradox of vagueness? In the paradoxical reasoning above, we move from a clear case of baldness, a statement with verity 1, to a clear case of non-baldness, a statement with verity 0. The argument is valid, and the principle of tolerance looks in good shape. Edgington's idea is that as we move along the sequence of persons each with a single more hair on their head than the previous one, we gradually move away from people who are clearly bald (verity 1), through the people where it is completely unclear whether they are bald (verity 0.5), to the people who are clearly not bald (verity 0). But at every point in the sequence, the relevant conditional that is an instance of the principle of tolerance is extremely plausible, because it has a verity just short of 1. For Edgington, the verity of a conditional just is the conditional verity of the consequent given the antecedent. And the conditional verity of x is bald, given that y is bald, is the value to be assigned to x is bald on the hypothetical decision to count y as bald, a decision that is not clearly wrong, given that y is a borderline case of baldness. By employing conditional verities, then, Edgington hopes to explain the plausibility of the instance of the principle of tolerance used in the paradoxical reasoning.

Edgington's approach to the paradox of vagueness is intriguing, but it raises many questions (as do all approaches to vagueness). One set of questions concerns the nature of verities themselves. Edgington is clear that verities are not degrees of truth and so are not intended to replace the classical bivalent truth-values. In Chapter 15, Nicholas Jones argues that Edgington is mistaken to think that verities and the classical truth-values are not in competition because classical semantics is incompatible with plausible principles concerning the relationship between the two approaches. Jones also casts doubt on Edgington's claim that verities are not in fact truth-values.

However we ultimately understand verities, Edgington's approach is motivated by what she takes to be analogies between the paradox of vagueness and the preface paradox. In Chapter 14 Alan Hájek argues that Edgington was correct to draw parallels between reasoning with uncertainty and reasoning with vague concepts. Hájek points to experiments in which subjects are taken along a series of coloured patches, where such subjects display so-called *reverse hysteresis* in their responses. In the experiment Hájek discusses, subjects are presented with a series of colour patches ranging from clearly blue, through bluey-green patches to those that are clearly green. What happens is that there are patches that subjects label as green when approaching from the blue-end of the range that they label as blue when approaching

from the green-end of the range. Hájek takes such judgments to be rational and argues that the best explanation of this is that this is a version of the Preface Paradox, the Progressive Preface Paradox.

Rosanna Keefe, whilst not wanting to deny some analogies between the preface paradox and the paradox of vagueness, argues in Chapter 13 that there are important disanalogies between reasoning with vague concepts and the preface paradox and that this constitutes a case against Edgington's treatment of vagueness. In particular, Keefe argues that whereas in the preface paradox we believe all of the premises individually, but not their conjunction or universally quantified form, in the case of the paradox of vagueness we believe both the individual premises and their conjunction and universally quantified form. Keefe, instead, argues that a supervaluationist treatment of vagueness is better equipped to take account of these facts.

1.3 The Paradox of Knowability

Edgington's final contribution to be discussed here, is her novel take on the so-called paradox of knowability. It is now widely known that a weak form of verificationism *classically* entails an absurdly strong form of verificationism, given certain seemingly minimal assumptions. In particular, reading 'Kp' as p is known by someone at some time or other, the weak form of verificationism:

Knowability $\forall p(p \rightarrow \Diamond Kp)$

entails the implausibly strong version of verificationism:

Known $\forall p(p \rightarrow Kp)$.

Proof:

- | | | |
|-----|---|--|
| 10. | $q \wedge \neg Kq$ | Assume there is an unknown truth for reductio; |
| 11. | $(q \wedge \neg Kq) \rightarrow \Diamond K(q \wedge \neg Kq)$ | Instance of Knowability; |
| 12. | $\Diamond K(q \wedge \neg Kq)$ | 10, 11, modus ponens; |
| 13. | $\Diamond(Kq \wedge K\neg Kq)$ | 12, Knowledge distributes over conjunctions; |
| 14. | $\Diamond(Kq \wedge \neg Kq)$ | 13, The Factivity of knowledge. |

But no contradiction is possible, contra (14), so our original assumption, (10), is false and there are no unknown truths. So, Knowability entails Known—that all truths are known. But since the latter is unacceptable—no one knows how many hairs were on my head twenty years ago—so is Knowability. Knowability, then, has to go.

As presented here, the paradox of Knowability (or Fitch's Paradox, or the Church-Fitch Paradox) is first taken as a proof that Knowability entails Known, and second, given that Known is false, that Knowability is also false. But where is the paradox in that? Rather, it seems that we should take the proof not as a *paradox* but rather as a *result* showing that there are certain structural limitations on knowledge (cf. Williamson (2000: 271.)) In particular, the proof shows that where q is an unknown truth, the fact that it is an unknown truth cannot be known. Stated this way, the falsity of Knowability seems obvious.

Nevertheless, Edgington claims '[t]hat truths which are [in principle] unknowable... should abound in the form of the most ubiquitous and mundane

facts, such that no one noticed a fly on the ceiling, or when this leaf fell from this tree, strikes many as paradoxical' (2010: 42). Paradoxical or not, Edgington argues 'that there is a sense in which one can know that, as things actually are, p and it is not known that p , but from a counterfactual perspective—as it were, from a modal distance' (2010: 42). Edgington's thought is that the possible situation of the knower need not be identical to the possible situation of the unknown truth. Rather than vindicating Knowability, what Edgington argues for is the following (where s and s^* range over possible situations):

E-Knowability $\forall p \forall s ((\text{in } s: p) \rightarrow \exists s^* (\text{in } s^*: K(\text{in } s: p)))$.

Williamson (2000, and Chapter 12, this volume) raises a challenge for Edgington's defence of E-Knowability, namely, how is the situation of the truth, s , specified? It cannot be specified by a subject of some other situation, s^* , by use of the phrase 'the actual situation', since this phrase will pick out s^* not s . More generally, there seems to be no way that the potential knower can pick out s via some causal referential chain. A subject in s^* can pick out s by description, though: s is the situation in which $p, q, r, \dots$. But as Williamson notes, this renders E-Knowability trivial, which was not Edgington's intention, since the consequent of E-Knowability is true simply of someone knowing that in the situation in which $p, q, r, \dots$ obtain, p is true. How, then, to allow that a subject in s^* can pick out s , without allowing that this is trivial? Edgington's approach is to invoke counterfactual conditionals, but Williamson argues at length that this approach does not work, pressing, amongst other objections, the Frege–Geach worry for Edgington's view that conditionals are not truth-apt.

References

- Edgington, D. (1995) On Conditionals, *Mind* 104: 235–329.
- Edgington, D. (1996) Vagueness by Degrees. In R. Keefe and P. (eds), *Vagueness: A Reader*. Cambridge, MA: MIT Press, pp. 294–316.
- Edgington, D. (1997) Truth, Objectivity, Counterfactuals and Gibbard, *Mind* 106: 107–16.
- Edgington, D. (2010) Possible Knowledge of Unknown Truth, *Synthese* 173: 41–52.
- Gibbard, A. (1981) Two Recent Theories of Conditionals. In W. L. Harper, R., Stalnaker, and C. T. Pearce (eds), *Ifs*. Dordrecht: Reidel.
- Lewis, D. (1976) Probabilities of Conditionals and Conditional Probabilities, *Philosophical Review*, 85: 297–315.
- McGee, V. (1985) A Counterexample to Modus Ponens, *Journal of Philosophy* 82: 462–71.
- Williamson, T. (2000) *Knowledge and Its Limits*. Oxford: Oxford University Press.

2

Philosophy and Me

Dorothy Edgington

My earliest philosophical thought—or the earliest I recall—occurred at the age of three or four. I had been taught, parrot-fashion, to give the right answer to questions like ‘What’s 3 and 5?’ This skill was being shown off by my grandparents (with whom we lived during the war) to visitors, who were suitably impressed. And I was very puzzled about *why* it was ‘clever’ to say 8 as opposed to 7—what made the one a good answer and the other bad. That turns out to be a hard question.

My next philosophical memories are not until the final year at school, when, as preparation for the General Paper of the Oxford and Cambridge entrance exams, we were given essay topics of a philosophical kind. I recall one on the justification of punishment; another on whether there could be time without change. I discovered that consulting large dictionaries often helped me to get started. That year I also read some Russell on the philosophy of science—I think it was *ABC of Relativity*—and dipped into Adam Smith’s *The Wealth of Nations*. But I had only the dimmest awareness of philosophy as a subject, and it never occurred to me that it was what I would do. Maths was my favourite subject at school, especially geometry, proofs giving great satisfaction.

Neither of my parents went to university, nor any other relatives, except for one second cousin, before me. Both my parents did well at Forfar Academy, the one secondary school in town, and got a string of Scottish Highers. Both regretted not going to university. So education mattered to them for their only child: from an early age it was taken for granted that I would go—indeed, to Cambridge (assumed to be the best), though that didn’t come about. What would happen beyond university was never thought about.

I was on track at eleven, at Bolton School (Girls’ Division). Then we moved to Lima, Peru, and for four years I had a somewhat spotty education at Colegio San Silvestre. There was no science, which bothered me. There was no Latin, which bothered my mother, as she knew it was required for Oxbridge. She ended up teaching, for a year, a small class herself, which allowed me to scrape through Latin O-level (aided by some guesswork from Spanish). After O-levels I had run out of schooling there, and was sent to St Leonards, in St Andrews, to catch up with science and do A-levels.

I was accepted by St Hilda’s College, Oxford, to read Engineering. This was a crazy choice, though it might have helped me get in: one other girl and I were the first to do

this at St Hilda's (a women's college—there were no mixed colleges for undergraduates then). Girton College, Cambridge, had the good sense to turn me down for the Mechanical Sciences Tripos. (It wasn't just that I was unsuited to the subject. In those days, Cambridge's entrance exams were a whole term after Oxford's, and by that time my mind had strayed far from academic matters.) So Oxford it was. The first year was disastrous academically. I wanted to change to Maths, but the Maths tutor didn't want me. (I was not working very well.) I was, however, allowed to change to PPE—Philosophy, Politics and Economics.

In my first term of Philosophy I had tutorials with a graduate student, Timothy Potts (later of the Philosophy Department at Leeds). We focused on Wittgenstein's *Tractatus*, and on Frege: a rather non-standard introduction to the subject, but one I found fascinating. I also enjoyed my tutorials in Economics with Margaret Paul, sister of Frank Ramsey, and wife of an Oxford philosopher, George Paul, who, sadly, died in a sailing accident during my second year.

There was little choice in PPE in those days: two compulsory papers in each of the three subjects, and two optional papers. Still feeling a mathematician *manqué*, my options were Formal Logic—the relatively advanced special logic paper, taken by only three people in the University in my year; and the statistics paper in Economics. I was taught logic mainly by E. J. Lemmon, before he left for California where he also, alas, met an early death, while mountaineering. I enjoyed PPE, though never knew whether I was doing well or badly. In Philosophy, J. L. Austin, though recently deceased, was still very much alive, at least with my tutors, and I found that the pickier I was about what the words in the essay question might mean, the more I seemed to please them.

Staying at university after graduation hadn't occurred to me as a possibility until, via a friend, I got some casual research-assistant work for an American doing a doctorate in economics at Nuffield College. That introduced me to the graduate-student scene, and the assumption that if you were good enough, that is what you did. By what seemed like a fluke, my Finals result meant that I was good enough, so I went to Nuffield to do a B. Phil. in Economics. (Plan B was the Statistical Office in the Civil Service, plan C a job with IBM.)

Wrong subject yet again! This time it wasn't a disaster, and indeed that first year after graduation I learned a lot about probability, which became an abiding interest. But as the year progressed I found myself increasingly consulting the Philosophy lecture list, especially for Logic, and attending those classes with more enthusiasm than those in Economics. Towards the end of that year I went to see my old philosophy tutor, Sybil Wolfram, to raise the possibility of changing to Philosophy. She swiftly arranged for me to see Gilbert Ryle at Magdalen. Ryle treated the proposed change of subject as a *fait accompli*: he wrote a letter there and then, in appalling handwriting, to the funding authorities, insisting that I should be funded for the next two years for the B. Phil. in Philosophy despite having already had one year's funding for the B. Phil. in Economics. He handed the letter to me and told me to 'get it typed' and send the original with the typescript. Of course I had to type it myself, but it worked out.

Both my false starts were, I think, the result of a feeling that I should do something 'useful'. And in 1964 Economics did seem useful and important, with Harold

Wilson's 'white heat of the technological revolution' on the horizon. After Wilson's election, Oxford economists went back and forth to London a great deal, advising the government. One consequence was that we graduate students were in great demand as tutors, and I got plenty of teaching experience that year.

That summer, 1965, I married John Edgington, whom I had met just over a year before, a few months before Finals. John had just completed a Cambridge Ph. D. in Physics (though most of his research was done at Harwell and he had lived in Oxford); and he got a lectureship at Queen Mary College in London. We continued to live in Oxford, renting a lovely cottage, Grist Cottage in Iffley, from people who had bought it for their retirement a few years on.

Most of my work in the next two years was with Michael Dummett—a great privilege. There was a historical paper, 'The Authorities for the Rise of Mathematical Logic', mainly on Frege, Russell, Hilbert, and Brouwer; and another advanced logic paper. Dummett was wonderfully illuminating, and I never ceased to be in awe of him. The Dummettian line of thought connecting the theory of meaning, classical logic, and realism would often crop up. I found this fascinating, and later felt that those two years gave me a head start in understanding Dummett's anti-realist challenge. My third subject was Metaphysics and the Theory of Knowledge, and I don't recall doing much work for that. And I wrote a thesis on the concept of probability. Ramsey and de Finetti were my heroes. Carnap, Keynes, and von Mises were criticized. (I had not yet come across propensity theories of objective chance.) A few suggestions were made at the end that there might be more constraints on rationality than the subjectivists allowed, such as ascribing the same probability to events between which one saw no relevant difference. I worked on the thesis during my first term with William Kneale—surrounded by packing cases because he was about to retire. After that I had no supervision on the thesis. It was written in a great rush in the last available few weeks, and mostly typed up by John. (From a conversation years later I learned that at just that time, under just the same pressure, at Harvard, David Lewis was typing up his wife Steffi's master's thesis, and Steffi's memories of the tinkle of the typewriter, interrupted by expletives caused by typos, were just like mine.)

It was not the practice, then, in philosophy at Oxford to continue to a doctorate after the B. Phil. Rather, one tried to get a job. Our son Alec was born a few months after I finished the B. Phil. I hadn't applied for jobs, because there was a plan afoot for us to go to Nigeria for a year, to the University of Ife—a special scheme to upgrade the teaching of physics with a rota of visitors (and I think I was lined up to teach some economics). Indeed before he was born Alec had a plane ticket in the name of 'Inf Edgington'. However the Biafra war made us decide that this venture was unwise, especially with a newborn child. During the next year I did some tutoring and examining for economics A-level. (It mattered to me to be self-supporting. Beyond that I was not very ambitious.) I did then apply for a few jobs, and was eventually offered a lectureship in the Philosophy Department of Birkbeck College, London. I knew a longer gap would make it harder, if not impossible, to get a university job, let alone one in the same city as John's, so I had to accept—even if it was a bit scary, as I was expecting another child. It was a weighty decision, but an irresistible one.

We moved to London in the summer of 1968. I felt we must live near my work, and so we rented a flat in Mecklenburgh Square, Bloomsbury—the rent and rates

(i.e. council tax) together matched exactly my net salary. Just over a year later we moved to a larger—and cheaper because shabbier—flat in the same square, in which we live to this day.

Birkbeck was the college of London University which catered for mature, part-time students. Most were older, wiser, more knowledgeable, and more confident than I was. Most of my lectures required learning from scratch. One of my lecture courses was on the history of political philosophy! Our daughter Fiona was born in March. Maternity leave had not been invented. Lectures went on well into the summer, the third term running from late April to early July. It was a 'make-or-break' sort of year. Later, the Kingsway Crèche—London's oldest day nursery—and the larger flat allowing for live-in help, made life a little easier.

I did feel a bit out on a limb, philosophically, at Birkbeck to begin with, and often felt I didn't understand what was going on in discussion. Logic (which I taught) was not held in high regard. David Hamlyn, our Head of Department, in his introductory talk to new students, giving thumbnail sketches of his colleagues, liked to say 'Dorothy does sums'. He would then add 'If you can't do the sums, *don't worry!* They're really not important!'

But soon the subject moved more in my direction—most notably with the 'Davidsonic boom'. I was never a card-carrying Davidsonian, finding it implausible that the structure of natural language could be captured in the structure of first-order extensional logic. But it did mean that a certain amount of logical machinery—satisfaction of predicates by sequents, etc.—was required by anyone who wanted to understand this project, and logic was more in vogue once more. New colleagues, Mark Platts and later Ian McFetridge, were more congenial philosophically.

It wasn't just Davidson. Much more was happening in the early 1970s that engaged me. There was considerable interest in Richard Montague's possible-world semantics. (Hans Kamp, who had been a student of Montague's, was at UCL and then at Bedford College.) There was David Lewis's and Robert Stalnaker's work on counterfactuals, and Dummett's magnificent *Frege: The Philosophy of Language*. Saul Kripke's lectures, *Naming and Necessity*, circulated in booklet form. No other philosophical work have I found so riveting. It was a quite exceptional period.

1974–5 was spent in Vancouver at the University of British Columbia, John to work at the new cyclotron, and I had a visiting position in the Philosophy Department. It was a good year, and I benefited greatly from two seminars given by Jonathan Bennett, one on a draft of his book *Linguistic Behaviour*, the other on his critical notice of David Lewis's *Counterfactuals* (*Canadian Journal of Philosophy*, 1974). Jonathan found me a useful critic: I had come from a setting where the themes of both seminars were much under discussion.

The counterfactuals seminar is a vivid memory. The classroom, like many at UBC, had stunning views over sea, islands, and mountains. The time, unusually, was late, around sunset. If you arrived early, you didn't turn on the light, you didn't speak, you just looked. Jonathan would then arrive with a huge urn of coffee, to keep us going late into the evening.

Jonathan's critical notice was in press. The seminar was based largely on it, but also conducted in the spirit of Goodman vs. Lewis. Jonathan tended to favour Goodman (he had some interesting ideas for circumventing the charge of circularity concerning

cotenable). Throughout, I defended Lewis. I did so with some success. Then, early in the final class, came the lethal objection. All eyes were on me, expectantly. I said nothing. After a long pause, Jonathan said 'I think we can finish there.' Curtain down.

The lethal objection was of a piece with Kit Fine's better-known and perhaps more memorable example, 'If Nixon had pressed the button, there would have been a nuclear holocaust' (from Fine's critical notice of Lewis's book in *Mind*, 1975). By ordinary standards of similarity, the most similar worlds to the actual world are not always the worlds that would have come about, had the antecedent been true. This was, of course, before Lewis replied that it is not ordinary standards of similarity which are at issue. But the examples convinced me that similarity to the actual world is simply not the right notion to delimit the relevant worlds. Ironically, in later work, Bennett gravitated towards a Lewisian theory, and I gravitated away.

I also worked on probability that year. I read Jon Dorling's excellent review article on Mary Hesse's *The Structure of Science* (*British Journal for the Philosophy of Science*, 1975). I found a contradiction in his analysis of how Bayesians could secure convergence of opinion when evidence, on average, went in the same direction. We corresponded about this for a bit, and jointly published a note which was in effect a correction to that part of his review (*BJPS*, 1976).

Yet another seminar at UBC motivated me, given by Ed Levy, on quantum logic. I never did get a very good understanding of the issues in quantum theory. But I came to see how close the connection was between classical logic and standard probability theory. For instance, using just probability theory one can prove that the probability of $\neg\neg A$ must equal the probability of A , the probability of $A \vee \neg A$ is always 1, $p((A \otimes B) \vee (A \otimes C)) = p(A \otimes (B \vee C))$, etc. A non-standard logic requires a non-standard probability theory; and, assuming this can be developed, the question arises whether the non-standard probability theory is plausible, interesting, useful, and powerful. If not, there may be arguments for classical logic from probability theory. I worked on this in the late 1970s, and it led to a paper on Dummett's challenge to realism, 'Meaning, Bivalence and Realism' (*Proceedings of the Aristotelian Society*, 1980–81). I was motivated by two remarks of Dummett's in the long and substantive Preface to *Truth and Other Engimas* (1978): first, we should not assume that we can 'simply transfer to empirical statements the intuitionist account of mathematical ones, since obviously there (are) great dissimilarities between them' (p. xxix); second, 'it is misleading to concentrate too heavily, as I have usually done, on a form of anti-realist theory of meaning in which the meaning of a statement is given in terms of what conclusively verifies it; often such conclusive verification is not to be had' (p. xxxviii). In the paper I argued against the intuitionist accounts of 'or' and 'not' in non-mathematical contexts; and I showed that once we have admitted non-conclusive justifications for asserting, accounts can be given of negation and disjunction which meet Dummett's constraints, but do not cast any doubts on the law of excluded middle and bivalence. Probability theory was in the background, but certainly not in the foreground: I kept my assumptions as weak and as intuitive as possible. I find, on re-reading, that I still rather like that paper!

Dummett was in the chair when I gave the paper at the Aristotelian Society. While the outcome was not one that he welcomed, he did say he was pleased that I was addressing his challenge head-on; and he had, as always, many insightful remarks to make.

Act II on conditionals begins, back in London, late 1975 or early 1976, when David Hamlyn, then Editor of *Mind*, comes into my office with two books for me to review. One of them is Ernest Adams, *The Logic of Conditionals*. Reading it is a revelation. Two of my areas of interest, probability and conditionals, join up. Conditional judgements are often uncertain. They are assessed by conditional probability. Why hadn't I thought of that?

(For what it is worth, before reading Adams, on indicative conditionals, I had tried to persuade students that there were strong reasons to accept the truth-functional account, although there were also strong reasons against it; and, rather half-heartedly, I had invoked Grice to try to mitigate the case against. On counterfactuals, although I was initially very taken by Lewis, Bennett had persuaded me that similarity is not the right notion, and I had already thought that something like 'the most probable worlds' would do better.)

The cognoscenti—Lewis, Stalnaker, Jackson—had been familiar with Adams's articles from the mid-1960s, but I had not. Of course there was Ramsey's suggestion long before him (1929) but no one had paid much attention to it, before Adams. At the heart of Adams's work is a probabilistic consequence of the notion of a valid—necessarily truth-preserving—argument. Consider such an argument. Suppose you think, but are not sure, that the premises are true. What should your attitude be to the conclusion? Call the uncertainty of a proposition one minus its probability. Valid arguments have the property that the uncertainty of the conclusion cannot exceed the sum of the uncertainties of the premises. In that sense, valid arguments preserve probability: there can be no more uncertainty in the conclusion than there is distributed among the premises. This is easily proved. It is quite an intuitive result, but not so intuitive as to be obvious without proof. It explains the extent to which one can rely on valid arguments in uncertain contexts. And it explains what goes wrong in the lottery paradox and the paradox of the preface: a very large number of premises, though each very close to certain, can yield a certainly false conclusion. Now Adams's initial attitude was: we haven't been able to find satisfactory truth conditions for conditionals, but we have a good idea of how to assess them probabilistically. So let us call the valid arguments involving conditionals those which satisfy the above probabilistic constraint: they preserve, in his sense, probability or conditional probability. He then shows how, on this conception, 'If A , B ' does not follow from $\neg A$, and does not follow from B . More surprisingly (at the time), transitivity, contraposition, and strengthening of the antecedent fail.

Adams's work had greatly influenced Stalnaker, who, with the new and powerful tools of possible-worlds semantics, aimed at truth conditions for conditionals such that the probability of their truth is the conditional probability of consequent given antecedent. Alas, it could not be done. Lewis's proof to that effect (1976) is the most famous. But Adams had his own proof, earlier. Indeed, the result was intuited by Ramsey in 1929: 'Many sentences express cognitive attitudes without being propositions. . . . This is even true of the ordinary hypothetical.' It was also realized by de Finetti, the other founder of subjective probability theory, in 1935, when he developed his own theory of conditionals construed as judgements of conditional probability. And I have found that the result causes less surprise amongst probability theorists than it does amongst philosophers: a conditional probability is not the

probability that something is the case, *simpliciter*, but the probability that something, *B*, is the case *under the assumption that* something else, *A* is the case. The reason we haven't been able to find satisfactory truth conditions for conditionals is that they don't have any: they are not to be thought of as propositions, true or false as the case may be.

All this took a lot of getting to the bottom of. Perhaps I was eased into accepting the lack of truth conditions by having previously reviewed J. L. Mackie's *Truth, Probability and Paradox*, where he defends a suppositional view of conditionals and denies that they are straightforwardly true or false. I didn't much like the view at the time, and didn't think Mackie had adequate reasons for the denial, but it was at least something I had come across and thought about. I worked a lot on trying to get the arguments against truth conditions as simple and intuitively compelling as I could.

Leaping ahead nearly a decade, I gave a course on conditionals at the Instituto de Investigaciones Filosóficas in Mexico City, in the summer of 1985. My friend and colleague from Birkbeck, Mark Platts, had gone to work there. I had already visited him. I had surprised his colleagues there, and had surprised myself, by finding that I could still speak the Spanish I learned in Peru as a child. I gave the course in Spanish, which was great fun. In the audience was a recently arrived philosopher from Argentina, Raúl Orayen. Having each come about 6000 miles from opposite directions, it was sometimes uncanny how similarly we thought about topics in philosophy and logic. Raúl wanted a paper on this material for the Mexican journal *Crítica*. Almost without exception, publications need to be squeezed out of me. Raúl pestered a great deal, and 'Do Conditionals Have Truth Conditions?' appeared in *Crítica*, 1986. A year or two later, back in London, in my role as Honorary Secretary and Editor of the Aristotelian Society, I was corresponding with Frank Jackson about his contribution to a Joint Session, and enclosed a copy of the conditionals paper. He selected it for the collection on conditionals he was editing for *Oxford Readings in Philosophy* (1991) and so the paper had a wider readership than I had thought it would. (I did mess things up a bit by adding to the original version a few paragraphs towards the end, about Stalnaker's 'Indicative Conditionals', and making an error (about a rain dance). Others have convinced me that there is nevertheless a good point to be salvaged from the error.)

I was, from the start, very taken with Adams's ideas about counterfactuals, but we'll leave that topic until later. I think it was in the late 1970s that I was first struck by an analogy between the lottery paradox and the sorites paradox. The message of the former, as I said above, is that a sufficiently large number of premises, each very close to certain, can lead, by valid reasoning, to a certainly false conclusion, because the uncertainties of each premise can mount up. Similarly in the case of vagueness, I came to believe, a large number of premises each very close to clearly true can lead one by valid reasoning to a clearly false conclusion. Indeed I argue, somewhat unorthodoxly, that degrees of closeness to clear cases, idealized so as to be represented by numbers, have a probabilistic structure, and so yet another application for probability theory is to the phenomenon of vagueness. My most worked out paper on that theme is 'Vagueness by Degrees' in Rosanna Keefe and Peter Smith, *Vagueness: A Reader* (1996).

It must have been in the early 1980s when, working on Dummett, I came across Fitch's paradox: the argument from 'all truths are possible objects of knowledge' to 'all truths are known'. For suppose there is a truth, p , which is never known to be true. Then there is a truth, ' p and it is never known that p ', which it is impossible to know. It occurred to me—suddenly, one evening, on my bicycle—that there was a way round this argument: the 'world' of the knowledge does not have to be the 'world' of the truth. Just as we can have knowledge of other possible situations, so, in other possible situations, we can have knowledge of other possible situations, some of which are actual situations. Just as I can know that if I hadn't noticed, p would never have been known, so, when I didn't notice, in the possible situation in which I did, I can think of the actual situation and know that in it, p is true and unknown. All this needed a lot of making good, but trying to do so was interesting and fun. Three brief comments: first, I use 'knowledge' for simplicity, but it would be safer to state the whole argument in terms of reasonable belief. Second, the modal entities of which we have knowledge should not be thought of as fully specific fine-grained worlds, replete with every detail, but as coarse-grained possibilities, like the possibility that it will rain in London tomorrow—the sort of possibilities we refer to, talk, think, and reason about. There is much to be said, independently of Fitch's argument, for taking these as basic. And third, one needs to say something about what makes the possible knowledge latch on to the actual situation, rather than some other situation in which p .

In 1985, preparing a Joint Session paper for a symposium with Anthony Appiah on Dummettian themes, I found I needed to refer to this phenomenon, so I submitted my 'Paradox of Knowability' paper to *Mind*, where it was published in the same year. Two critical responses by Timothy Williamson appeared promptly, and were later reworked in his book *Knowledge and Its Limits*. Pressed to respond to his criticisms then, I tried to do so in *Synthese*, 2010, where I tried to do a better job of saying what makes a merely possible piece of knowledge refer to the actual situation. The debate continues.

I was not motivated by an urge to defend antirealism, although this argument did seem to me to defeat it too easily. It was rather an engaging project, trying to make sense of an 'outside view looking in', rather than an 'inside view looking out', at modal reality, to put it metaphorically.

In 1986 I spent the autumn term as a Visiting Fellow at Princeton, where I enjoyed some great seminars: David Lewis on modality, Saul Kripke on identity and Dick Jeffrey on probability. Bas van Fraassen was on leave, but back in town occasionally, and I had some good conversations with him about belief revision. (The only downside to the visit was a broken jaw and ruined teeth as the result of a bicycle accident.)

I made two visits to Prague in the 1980s, giving underground lectures, under the auspices of the Jan Hus Foundation. The first were part of a course on Frege and Russell, the second part of a course on Kant (the Aesthetic, and the Transcendental Deduction)—all ideologically dangerous in the eyes of the regime. The audiences consisted of highly intelligent and cultured street-sweepers and the like. A lecture went like this: I had a script, and so did the organizer. I would read a sentence. He would translate it. There often ensued a discussion or argument about the

translation, and I would then be consulted: is this, or that, a better rendering of what I said? There would follow a discussion about whether I was right. Only when the sentence had been understood and evaluated by everyone did we proceed to the next sentence. It took a long time, but never have I had such an attentive audience. The first lecture I gave on my return, back in London, I looked around in dismay and wondered what proportion of what I said was being taken in!

In 1988 I lectured again in Mexico, and went on to Buenos Aires to give some lectures at the Argentinian Society for Analytic Philosophy, SADAF. The Argentinian philosophers were very able and keen. They had recently suffered a period not unlike that in Prague: during the time of the Colonels, analytic philosophy was deemed dangerous and SADAF was banned, and had gone underground. That was a time of especially intense philosophical activity, they said. Now that democracy was restored, one of its members, Carlos Nino, was Minister for Justice.

SADAF occupied a pleasant town house a little north of the centre of Buenos Aires. The organization considered itself to be modelled on the Aristotelian Society (which I was running at the time). I was asked more than once, where in London is the Aristotelian Society? Their image was of an impressive building, so they were most disappointed to learn that we hired a room once a fortnight for our meetings, and apart from that, it was located nowhere! (I was able to arrange for them to receive free copies of the *Proceedings*.)

I taught at UBC in Vancouver again in 1990 and 1992, one semester each time. On the second occasion I was told, rather apologetically, that what was required was an introductory course on non-conclusive reasoning: two sections, each meeting three times a week, and doing weekly homework. This turned out to be most enjoyable—loads of probabilistic puzzles, and much besides. I have notes exceeding 100 pages. I found the experience of doing the same course twice over, to my surprise, a very good way of improving the course. Back in London, in what was still the federal degree, there were daytime lectures for UCL and King's students and evening lectures for Birkbeck students. I suggested that those who wished could meet their lecturing requirements by giving the same course, once in the day and once in the evening, rather than two separate courses. Many liked to do this, though some didn't, and it came to be known as the Edgington plan, and flourished for well over a decade. Alas, the federal degree is no more and there is little opportunity for co-operation between the colleges these days.

Also perhaps worth a comment is a method of distributing notes which I learned from Jonathan Bennett: I would give the lecture, then I would write up the notes—always a good lead in to preparing the next lecture—and they would be distributed at the next meeting, after the event. I did this later when teaching at Oxford (mainly logic and language, also epistemology) and most students liked it—certainly, I had requests for copies of the notes for years after I left Oxford.

In 1992 John had an overdue sabbatical coming up after an exceptionally long stint as Head of Department, I applied for and got a British Academy Research Readership, and we spent the academic year 1992–3 first at the University of Texas at Austin, and then at Berkeley. At Berkeley we were lucky to rent Donald Davidson's house, as high on the hill as the houses go. And at Berkeley I finally met Ernest Adams, and we worked together a lot—not only on conditionals, but also on the

prospects for a probabilistic degree-theoretic account of vagueness. I wrote some papers on vagueness that year.

The main product of the Research Readership was the long 'state of the art' article on conditionals Mark Sainsbury asked me to do for *Mind* (1995). Back in London after the sabbatical, my new colleague Scott Sturgeon was a big help when I was working on this, and became a good friend. Most of this paper I am still pleased with, but the last section on counterfactuals was rather a rush and not so much to my liking. This was the first time I published on counterfactuals, although I have done so since, and I had been thinking about them for a long time. In the first paragraph of my review of Adams's book, I wrote 'best of all, [a theory] in terms of which the notorious counterfactual conditional can easily be explained'. I hereby retract the word 'easily'! But I thought, and still think, Adams had good ideas about counterfactuals as well as about indicatives. Having become disenchanted with the notion of similarity, it had already occurred to me that probability would do better. To put it in terms close to Lewis's, we need a probability distribution over the relevant antecedent-worlds, and we need to consider the probability of the consequent in that distribution. That is a conditional probability. Evidence for this view is the seamless transition from countless forward-looking indicatives, to the counterfactual, when the antecedent proves false. 'It's about 90% likely that you will be cured if you have the operation' becomes later 'It's about 90% likely that you would have been cured if you had had the operation'. The difference between the two is that the counterfactuals are typically not assessed from your present epistemic position: typically, for counterfactuals, the question is, how likely *was* it that this *would* happen, given such-and-such? (There are complexities, however, about the standpoint from which we make these judgements.)

Why do we go in for such thinking? No doubt there are several reasons, but Adams highlighted an important one: they play a crucial role in empirical reasoning about what is the case. You make an observation. You ask: how likely *was* it that I *would* observe this, given various hypotheses; and reassess the probability of these hypotheses, in the light of the observation. That is, he connects our use of counterfactuals to ordinary Bayesian reasoning.

Which are the relevant worlds? Which actual facts do you hold constant and which do you abandon with the counterfactual supposition? These questions are faced by all theories. Lewis himself, specifying a 'default' answer, allows that this is a context-dependent matter. The best way of tackling it is focusing on what we want to use the judgement for. Even the famous Oswald-Kennedy counterfactual can be used in a non-standard way, just as a past-tense indicative: 'You had already got Oswald; why did you round up these other people in the crowd?' 'We weren't sure at the time that it was Oswald. If it hadn't been Oswald, it would have been one of these others who shot Kennedy.'

I also argue in the *Mind* paper that objectivity need not go out of the window with truth: we do our best to estimate the relevant objective conditional chances, when such there be, for forward-looking indicatives and counterfactuals.

In 1996 I was offered a position in Oxford—a five-year professorship to replace Christopher Peacocke, who held the Waynflete Chair, and who had been awarded a

five-year Leverhulme Research Professorship. I was glad to accept: I felt like a change of scene, and it felt like an honour. I had a Fellowship at University College.

There was a big difference between being a student at two non-traditional colleges, and being a Fellow of a very traditional college, which had its 750th anniversary during my stay. While the faculty life and the teaching all went well, I did find the college, in which I was living during the week, a rather alien environment. I used to say that the X90 bus, which took me weekly between Baker Street and the High, took on the aspect of a space ship which took me to another planet, with its own laws and rules, causally isolated from the world I knew (except for the X90 bus). From the domestic point of view, it was a relief to return to London in 2001. So when, in 2003, Christopher Peacocke having resigned to move to New York, I was eventually offered the job for real, I was in a quandary. But I am very glad that in the end I accepted. In the earlier period John was extremely busy in London (by now Vice Principal: he turned out to be too good at administration). By 2003 he had retired. We rented a house from the University in Walton Street. Magdalen was welcoming and beautiful. After the earlier stint, Oxford was less strange. The students were a delight to do philosophy with. Those were three good years. And it did seem amazing, having gone to see Ryle all these years back about changing to philosophy, to be occupying his Chair!

I was obliged (but willingly) to retire from Oxford, at sixty-five, in 2006. Birkbeck invited me back as a part-time Senior Research Fellow. Among the highlights since was a two-week summer school on conditionals at the Central European University, Budapest in 2009, my colleagues on the faculty being Alan Hájek, Angelika Kratzer, Barry Loewer, Bob Stalnaker, and Jason Stanley. One of those attending, Alberto Mura, invited me to a three-month visiting professorship in Sardinia in 2011, and that was a delightful time. Also in 2011, Lee Walters organized a conference in London to mark my seventieth birthday. It was lovely to see many old students and colleagues. Among the participants were David Over, my first doctoral student, whose work on the cognitive psychology of conditionals complements mine; Ruth Weintraub, another early doctoral student, from Israel; Ofra Magidor, one of the outstanding students of my last stint in Oxford (another star from that period, Sarah Moss, unfortunately could not come); and Nick Jones, then my most recent doctoral student, from Birkbeck. Of course I came in for criticism, as I shall in this book which arises out of that occasion, but that is how philosophy progresses. I am grateful to Lee in particular, for all the work that this has involved. Lee's thesis, at UCL, was on empty names, but he had a strong side-interest in conditionals, and I saw a great deal of him during his graduate-student years—especially in 2009 when, as well as the Budapest Conditionals Fest, there were conferences in Berlin and in Oslo where we both gave talks.

I have, of course, been highly selective in plotting a route through the chapter of accidents, many of them lucky, which is my time as a philosopher. I suppose what has gripped me most is reasoning—messy reasoning, involving uncertainty and vagueness. I am glad to have had work that is so engaging. The downside is that it is also so intractable. (For recreation, I like puzzles with solutions.) But most of the time it has seemed worth the struggle.

3

A Note on Conditionals and Restrictors

Daniel Rothschild

3.1 Introduction

Within linguistic semantics, it is near orthodoxy that the function of the word ‘if’ (in most cases) is to mark restrictions on quantification. Just as in the sentence ‘Every man smokes’, the common noun ‘man’ restricts the quantifier ‘every’, in the sentence ‘Usually, if it’s winter it’s cold’, ‘it’s winter’ acts as a restrictor on the situational quantifier ‘usually’. This view, originally due to Lewis (1975), has been greatly extended in work by Heim (1982) and, most notably, Kratzer (1978, 1981, 1986, 2012) into a rich theory of almost all uses of the word ‘if’. I call this the *restrictor view* of ‘if’.

Despite its linguistic prominence, this view of the word ‘if’ has played little role in the philosophical discussion of conditionals. Fairly recent philosophical surveys such as Bennett’s (2003) book-length introduction or Edgington’s (1995, 2008) review articles do not even mention the restrictor view. Stranger still, in his seminal work on conditionals and probability, Lewis (1976, 1986) does not discuss the restrictor view that he pioneered, despite the intimate relation noted by Kratzer (1978, 1986).¹ This chapter tries to fill in the gap left by these omissions.²

I make four main points. First, I argue that given the current state of affairs our best bet is to accept the ‘restrictor view’ and to assume that ‘if’ is not ambiguous, so that we should accept some variant of the full Heim/Kratzer account of conditionals. Second, I argue that the restrictor view is compatible with all major philosophical views of conditionals, if they are understood in the right way, namely as theories about the meaning of certain sentences that include ‘if’, rather than as theories about the meaning of the word ‘if’ itself. Third, I argue that the restrictor view undermines

* I am grateful to Dorothy Edgington, Justin Khoo, Dilip Ninan, Scott Sturgeon, Seth Yalcin for comments and discussion. I am particularly grateful to Angelika Kratzer and Lee Walters for detailed comments.

¹ It seems to me that Lewis must have thought ‘if’ was three-ways ambiguous: it acts as a pure restrictor under adverbs of quantification (Lewis, 1975); it is the material conditional in cases of indicative conditionals (Lewis, 1976); and it is a variably strict conditional in counterfactuals (Lewis, 1973).

² Many of the points made here expand on observations in Kratzer’s own work, unpublished lectures by von Fintel (1994), as well as Cozic and Égré (2010), and Rothschild (2013). Since this paper was drafted and put online in 2011, Kratzer’s view has been more prominently discussed in the philosophical literature.

an important argument from the probabilities of conditionals to a non-propositional view of conditionals (an argument which Lewis played a large role in developing). Fourth, I argue that consideration of embeddings of conditionals, while not decisive, provides some evidence for a combination of the restrictor view with the view that indicative conditionals express propositions.

Not all these points are completely novel, but I think together they paint an important picture of the current state of our understanding of conditionals, one which is not easily found elsewhere.

3.2 Conditionals and Semantic Theory

When linguists and philosophers discuss conditionals they are mostly talking about sentences that include the word ‘if’, such as these:

- (1) If a man comes in, he’ll be angry.
- (2) Usually, if a man comes in, he’ll be angry.
- (3) If a man comes in, he’ll probably be angry.

Philosophers often discuss rival theories of conditionals: the material conditional (e.g. Grice, 1967/1989; Jackson, 1987), the Stalnaker conditional (e.g. Stalnaker, 1968), the (related) strict and variably strict conditionals (e.g. Lewis, 1973; Ellis, 1978), the non-propositional theories (e.g. Adams, 1975; Edgington, 1995). In order to assess how these views relate to the restrictor view, we need to relate these theories to semantics generally.

For this reason I’ll say a bit here about the structure of semantic theory. Semantics aims at a systematic account of the meaning of sentences in terms of the meaning of their parts and how they are put together. This typically involves assigning meanings to words (lexical items) and specifying rules of semantic composition (i.e. rules that get you from syntactic structures with meaningful components to the meanings of the whole structures). In combination, then, we can assign meanings to entire sentences. In the case of a complete declarative sentence, a standard semantic theory will assign a proposition to it, in particular the proposition that the sentence expresses.³

Empirically this enterprise is constrained by what propositions sentences actually express, as revealed by such things as our truth-value judgments of sentences in different situations, our judgments of entailment and so on. Further constraints on the project come through the related concerns of simplicity and learnability. These concerns push for simple, clear rules of composition and simple, unambiguous meanings assigned to lexical items. Of course, there is ambiguity and complexity in language but we aim to build simple theories to capture these complexities. In addition, of course, the actual syntactic structure of sentences will constrain our theorizing as it is this structure that the composition rules need to work with.⁴

³ Of course doing so will often rely on using contextual information. I’m abstracting away from this here, as I don’t think it’s relevant to the particular points I am making.

⁴ Constraints like compositionality, which are motivated by concerns of simplicity and learnability, provide particularly sharp constraints on which theories are acceptable for given syntactic structures.

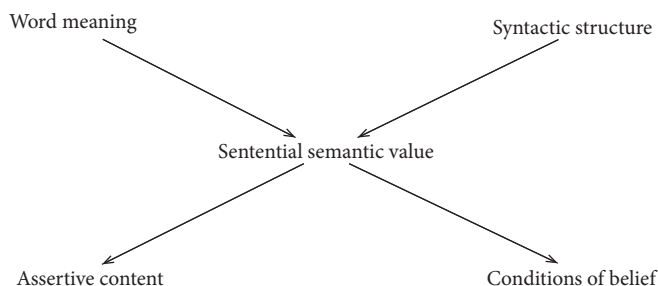


Figure 3.1. Structure of semantic theory for declaratives.

Semantics connects to the more personal-level notions of communication and belief mostly by way of the semantic values of entire sentences.⁵ Figure 3.1 shows the structure of the situation: assertive content and conditions of belief only connect to word-meaning via sentence meaning. The way in which sentential semantic values connect up to assertion and belief is mostly simple and familiar: If our semantic theory assigns a proposition p to a sentence S , then an assertive utterance of S is an assertion of p . Likewise, believing S is true amounts to believing p is true.⁶

It is important to note that even orthodox semantic theories do not always work by assigning propositions to sentences. Semantic theories typically do not assign propositions to 'wh'-questions—e.g. 'Who came?', 'Where is Kate?'—as their meaning. Rather the semantic values assigned to wh-questions tend to be *sets* of propositions or partitions of logical space (Hamblin, 1973; Karttunen, 1977; Groenendijk and Stokhof, 1984). That is because the speech act of questioning does not amount to the assertion of a proposition, but something more like a request for information.

In all cases, we implicitly or explicitly use *bridging* principles that connect up semantic values with the personal-level acts and states associated with the sentences (e.g. assertion and belief in the case of normal indicative sentences, asking and wondering in the case of questions). In the case of declarative sentences these bridging principles, evoked implicitly above, are trivial, i.e. if a sentence S has the semantic value p then an assertive utterance of S is an assertion of the proposition p .⁷ In Figure 3.1, the bridging principles are what connect the sentential semantic value to the assertive content and the conditions of belief; semantic theory, by contrast, takes us from word meaning and syntactic structure to sentential semantic value.

The point of this sketch of semantic theory is to illustrate the number of different levels at which claims about conditionals can be understood. I will argue here that the most charitable interpretation of philosophical theories of conditionals is as claims about (a) the semantic values of entire sentences that include conditionals, and (b) what it is to assert/know/wonder about those sentences, i.e. how we should

⁵ Of course, there are also the more elusive, sub-sentential speech-act notions of reference and predication that might constrain our semantic theorizing.

⁶ I am putting issues of context dependence aside here.

⁷ Except to the extent that semantic values of declarative sentences might be index-dependent as argued by Lewis (1980). See also Stanley (2002), Ninan (2010), and Rabern (2012).

understand the speech-acts and psychological states associated with conditional sentences. My main claim is negative: philosophical theories of conditionals should *not* be viewed as direct claims about the meaning of the word ‘if’ and the compositional rules that govern sentences with ‘if’. In terms of Figure 3.1 philosophical theories only cover the middle and the bottom sections of the picture. Thus, philosophers should not be seen as giving semantic theories of ‘if’ in the usual sense.

This claim may be surprising. After all, philosophical theories of conditionals tend to come as complete packages: theories of the meaning of the connective ‘if’, the meaning of entire sentences that include ‘if’, what is asserted by sentences that include ‘if’, and what it is to believe such sentences. Indeed, theories are often classified according to their view of the connective ‘if’: hook (the material conditional), the Stalnaker conditional, the strict conditional, etc. (as in Edgington, 2008).

My claim about how to best understand philosophical theories of conditionals does not, however, rest on the intentions of those propounding the theories. Rather it relies on the principle of charity: philosophical theories of conditionals are most plausible if understood at the higher level.

3.3 Conditionals and Adverbs of Quantification

Here I will sketch Lewis (1975), Kratzer (1978, 1981, 1986), and Heim’s (1982) view that ‘if’ is a device for marking the restriction of a quantifier. I believe this is one of the best established claims in semantic theory due to its simplicity and explanatory power. Lewis, in ‘Adverbs of quantification’, considered sentences like (4), where, intuitively, a conditional is embedded under an adverb of quantification.

- (4) Usually, if Mary is here, she is angry.

It seems reasonable to assume here that ‘usually’ functions as a quantifier over times or situations.⁸ In this case both ‘Mary is here’ and ‘she is angry’ will be true or false relative to different times or situations. This leaves the question of what the meaning of the conditional connective ‘if’ is in this case. We might think, as is standard in logic, that it is a connective that joins together the sentences ‘Mary is here’ and ‘she is angry’ to produce some complex sentence which itself is true relative to different situations. Lewis argued that this is not the right way to think about examples like (4). Rather, Lewis suggested, the entire ‘if’-clause, ‘if Mary is here’, acts as a restrictor on the quantification over times or situations. So we can paraphrase (4) as follows:⁹

- (5) Most situations in which Mary is here are situations in which she is angry.

⁸ In fact, Lewis rejects the *general* claim that adverbs of quantification are always situation quantifiers. He argues instead that they are unselective quantifiers that can quantify over any free variable. However, this aspect of Lewis’s theory is not generally accepted. In cases like (4), anyway, even Lewis would presumably think the right analysis has the ‘usually’ bind a time, event, or situation variable. I will assume in this note, following von Stechow (1994, 1995), that adverbs of quantification always bind situation variables, though nothing essential rides on this assumption.

⁹ I am putting aside here the various difficulties in counting situations which affect the interpretation of (5). See von Stechow (1997/2005); Kratzer (2011).

Thus, the function of ‘if’ in sentences like (4) is simply to mark the fact that ‘Mary is here’ is a *restrictor* of the situational quantifier ‘usually’. More explicitly: we think of all situational quantifiers, such as ‘usually’, as binary quantifiers that take both a restrictor and a matrix predicate.¹⁰ The semantic contribution of ‘if’ is to mark the fact that the material following it serves as part of the restrictor. The other material, what we traditionally call the consequent, goes into the matrix.

We can write a binary quantifier Q acting on the restrictor φ and the matrix ψ as $Q[\varphi][\psi]$. Thus (4) has the schematic form in (6):

- (6) Usually[Mary is here][Mary is angry].

To my knowledge there is no serious rival theory to Lewis’s account of the role of ‘if’-clauses under adverbs of quantification.¹¹ As Lewis points out, it follows from well-known results on binary quantification that no truth-functional conditional connective can predict the same truth conditions.¹² There are also no extant, plausible non-truth-functional accounts of conditional propositions that capture this equivalence.¹³

A technical note: As we will see, the restrictor view admits of many implementations within specific semantic frameworks. One possible view, adopted by Heim (1982) and Kratzer (1978) gives a syntactic spin to the view. Quantifiers, generally, are seen as having two arguments. Whether a given piece of syntactic material occupies one argument place or the other is a syntactic matter, and ‘if’ serves a syntactic marker that what follows it is in the restrictor argument place. On this syntactic spin ‘if’ has no semantic value whatsoever, it merely serves to mark a syntactic place for the material after it. (I give a simple version of this syntactic story in the first appendix.) This is by no means the only view we can have and it does not fit well with current syntactic theory. We can also think that ‘if’ takes the material inside it and returns a function that modifies quantifiers by restricting them with that material. In this case, ‘if’ has a very specific meaning: it takes as input a sentence and returns something that can modify quantifiers or their parameters (Kratzer, 1981, 2012). This view fits well with the idea that ‘if’-clauses are adverbial phrases (Geis, 1970). There are other possible views which we can think of as versions of the restrictor view, such as Belnap’s (1970) trivalent view, which I will discuss later. Which view you want will depend, mostly, on a lot of detailed questions about your overall syntactic and semantic framework, and I don’t think those questions much affect my discussion here.

¹⁰ Unary quantifiers like $\forall x$ and $\exists x$ take a single open-formula, e.g. Fx . A binary quantifier, such as **most**_x takes two open-formulas, e.g. Fx and Gx , one of which is called the restrictor the other a matrix. For example, in the sentence ‘Most men are tall’, ‘man’ is the restrictor predicate and ‘is tall’ is the matrix predicate. See, e.g., Barwise and Cooper (1981) for further discussion.

¹¹ I consider Belnap’s (1970) trivalent account of conditionals under quantifiers as one particular *implementation* of the restrictor view. Of course, if one does not have such a catholic view, then this would be a ‘rival’ to the restrictor view. There seems little point in quibbling about this issue.

¹² See Barwise and Cooper (1981) for discussion of this result which was originally proved by David Kaplan in 1965. Of course, this result only holds in a bivalent context, hence the possibility for Belnap’s trivalent semantics of conditionals.

¹³ Even elaborate dynamic accounts such as Gillies (2010) are not obviously capable of treating adverbs of quantification, as Khoo (2011) argues.

3.4 Uniformity

Semantic theories aim to be simple. Thus, in general, we should try to posit a non-ambiguous, simple meaning for ‘if’, as ambiguities add to the complexity of our semantic theories. Given that the restrictor analysis seems necessary for examples of conditionals under adverbs of quantification like (4), all else equal, we should apply it as widely as possible.

Kratzer (1981, 1986) and Heim (1982) showed the analysis can be expanded very widely. Kratzer noted that the analysis works well for conditionals that are embedded under various modal constructions. For instance, the analysis is easily extended to this set of examples:

- (7) a. Necessarily, if Mary is here, she is angry.
 Probably, if Mary is here, she is angry.
 It’s likely that if Mary is here, she is angry.
 If Mary is here, she must be angry.

In all these cases it is natural to see the modals ‘probably’, ‘necessarily’, ‘it is likely’, and ‘must’ as quantifiers over possible worlds that are restricted by the ‘if’ clause.¹⁴ So, if we treat modals as binary quantifiers we can give the basic semantic structure of the sentences in (7) as in (8).^{15,16}

- (8) a. necessarily [Mary is here][Mary is angry]
 probably [Mary is here][Mary is angry]
 it’s likely [Mary is here][Mary is angry]
 must [Mary is here][Mary is angry]

Kratzer, more controversially, argued that even in conditionals without explicit modal operators there are implicit modal operators. In particular, Kratzer argues that a bare indicative conditional—i.e. a conditional sentence without a higher modal operator, such as (9-a)—includes a silent necessity operator similar to ‘must’ or ‘necessarily’. Thus, the semantic structure of (9-a) can be represented in (9-b):

- (9) a. If Mary is here, she is angry.
 b. Must [Mary is here] [She is angry]

While the syntax and motivation of this view is novel, it follows in a long tradition of viewing bare conditionals as expressing a form of conditional necessity. So, in terms of its sentential semantics, it is a familiar view of bare indicative conditionals.

¹⁴ Of course, ‘probably’ isn’t a normal quantifier over worlds, but rather one that depends on a probability measure over the worlds (see Yalcin, 2010, for discussion).

¹⁵ I am assuming here that we assign suitable semantic values to the modal quantifiers, e.g. ‘necessary’ is a binary quantifier taking two sentences, a restrictor and a matrix, such that ‘necessary’[restrictor][matrix] is true iff in every world in which the restrictor is true, the matrix is true.

¹⁶ I’m only using the idea that modals are binary quantifiers as one illustrative way of doing the syntax and semantics here; as I mentioned in the previous section, we could instead treat the modals as unary operators that are modified by ‘if’-clauses. The relevant point here is that the ‘if’-clause has a semantic value of its own that serves to restrict the modal operator, rather than combining directly with the consequent.

I should note that this is not the only theoretical option for treating conditionals without overt quantifiers. Another kind of view assumes that the conditional expression (i.e. a bare conditional with both antecedent and consequent, like (9-a)) has some semantic value *X*. When a binary quantifier, like ‘necessarily’, applies to *X* we get restricted quantification. Our semantics, though, also assigns an interpretation to *X* of some sort when there is no syntactically present quantifier. An instance of this kind of view, perhaps the most minimal implementation, is Belnap’s (1970) trivalent view, which I turn to in the next section.¹⁷ What differentiates this type of view from the traditional Kratzer/Lewis view is that it assigns a single syntactic entity to the conditional expression ‘if Mary is here, she is angry’, rather than splitting it into two distinct entities. For this reason this view is not compatible with the syntactic construal of the restrictor hypothesis: ‘if’ has a semantic value here, it doesn’t just mark a syntactic place. However, when bare conditionals are embedded under quantifiers the results are equivalent to the syntactic construal: the antecedent restricts the quantifier.

So it is feasible (in more than one way) to give a unified analysis of bare conditionals and conditionals under adverbs of quantification and modal operators. Methodological considerations strongly support a unified analysis.

3.5 Restrictor-based Theories

I argued above that the most promising account of the meaning of the word ‘if’ is that it serves to mark the material after it as restricting some sort of quantification. This view usually does not even get mentioned in standard philosophical discussions of conditionals (e.g. Bennett, 2003; Edgington, 1995). There is a good reason for this: philosophical views focus on unembedded conditionals without explicit modal operators:

- (10) If Mary is here, she is angry.

Bare conditionals are obviously the toughest cases for the restrictor analysis since there is no explicit operator for the ‘if’-clause to restrict. When focusing on examples like (10) the restrictor analysis is unintuitive. Nonetheless, as I argued above, the restrictor analysis is the only game in town for examples like (4), and it is both unintuitive and bad methodologically to treat the ‘if’ in (10) as different from the ‘if’ in (4). So philosophers, if they are seriously interested in the word ‘if’, should presumably adopt as one of the most plausible hypotheses that ‘if’ in (10) is doing what it is doing in cases with adverbs of quantification. Since they generally do not do this, we might be tempted to dismiss philosophical theories as implausible.

Instead, I suggest we understand the major philosophical theories of conditionals as views about the semantic value of entire sentences with conditionals and views about which speech-acts are associated with such sentences. When viewed in this way

¹⁷ Lewis (1975) discusses this as a possible treatment of adverbs of quantification. In the context of probability operators the view can be found originally in de Finetti (1936). See Huitink (2008); Rothschild (forthcoming a) for further discussion of the trivalent view from a linguistic perspective.

the restrictor view of ‘if’ poses no challenge to the philosophical theories, since they are, as such, compatible with the restrictor view.¹⁸

3.6 Strict Conditional

Kratzer and Heim’s view is that a bare indicative conditional such as (10) contains an implicit modal operator. So the logical form of (10) is something more like this:

- (11) necessarily [Mary is here][she is angry]

As I noted above, this amounts to the view that bare conditionals express conditional necessity: in all worlds in which Mary is here, she is angry.¹⁹ So the restrictor view is obviously compatible with the strict conditional view, once we understand that as a view about bare conditional sentences rather than a view about the connective ‘if’.

3.7 Material Conditional

We can get the material conditional as a sort of limiting case of the strict conditional. Simply assume the necessity modal only quantifies over worlds that are actual. Since there is only one, the one question is whether the consequent is true at that world if the antecedent is. If the antecedent is not true at the actual world, the quantification is vacuous and so the sentence is true. Thus we get the truth-conditions of the material conditional. Of course, it is widely acknowledged that the material-conditional view is hopelessly implausible as a semantics for the conditional: it simply does not account for much basic data about truth-value judgments of conditionals.²⁰ Nonetheless, it is useful to see that the material conditional view, as a view about the semantics of bare conditional sentences, is not ruled out by the restrictor view alone.

3.8 Stalnaker/Lewis Conditionals

Stalnaker and Lewis propose that to evaluate conditionals one needs to look at the ‘closest’ possible worlds in which the antecedent is true.²¹ Kratzer implements the variably strict semantics for conditionals within her general approach to modality: all modals introduce both a base (a set of worlds) and an ordering on those worlds.

¹⁸ Kratzer, herself, made this point with respect to most of the propositional views of conditionals; my main contribution here is to extend this point to non-propositional views of conditionals.

¹⁹ It is widely recognized that this view is only plausible if we view the necessity operator as quantifying over a sharply restricted set of worlds (rather than, say, all metaphysically or physically possible worlds). However, in natural language semantics it is normal to think that all quantifiers are sharply restricted by context, so this does not seem like a problematic discussion.

²⁰ I think the assumption that all indicative conditionals with false antecedents are true flies in the face of many of our truth-value intuitions, and no amount of pragmatics can explain this fact away.

²¹ Lewis, of course, only thought this view should be used for counterfactual conditionals, while Stalnaker thought it should apply to all uses of conditionals.

'If'-clauses are still simply restrictors, but the modals do the work of ensuring that the worlds where the consequent is evaluated are the 'closest' worlds.²²

3.9 Non-propositional

The compatibility of the view of 'if'-clauses as restrictors with the major propositional views of conditionals was emphasized by Kratzer. However, there is little discussion in the semantics literature of the relationship of non-propositional views of indicative conditionals to the restrictor view, despite the prominence of non-propositional views in the philosophical literature. There are a variety of non-propositional views that accord with the restrictor hypothesis. I will discuss two such views here: a trivalent view and a view that combines a non-propositional semantics for epistemic modals with the restrictor view of 'if'. Before doing so, I will make some general comments about non-propositional views of conditionals.

It is often at least implicitly assumed that non-propositional views of conditionals are premised upon a rejection of the project of formal semantics at least insofar as it extends to include conditionals. This is a mistake.

It is true that the standard assumption underpinning almost all work in semantics is that when the semantic value of a sentence is a proposition then an assertion of the sentence is an assertion of that proposition and belief in the sentence is belief in that proposition. If sentences do not have propositions as semantic values, however, that does not mean we cannot do semantics. What we need, in this case, is new bridging principles connecting non-propositional semantic values with assertion and belief. The semantics in combination with these principles then makes predictions about what people can do with the relevant sentences. Even orthodox semantic views sometimes use non-propositional semantic values for complete sentences and, associated with them, non-standard bridging principles. As I mentioned earlier, a salient example where orthodox theories need such non-standard bridging principles is the semantics of questions. Groenendijk and Stokhof (1984), for instance, assign questions partitions of logic space as their semantic value. Asking a question is inquiring which cell of the partition the actual world lies in. Wondering about a question is wondering which cell the actual world is in. With this class of semantic values and bridging principles we can then judge whether certain assignments of semantic values are reasonable or not.

Non-truth-conditional programs about conditionals are not generally put forward as full-fledged semantic theories with explicit semantic values and bridging principles. This is often taken (by semanticists and linguistically-inclined philosophers of language) as an implicit rejection of the methodology of semantics. This does not seem fair to me. As I understand Edgington's (1995, 2008) view, she is not committed to a particular account of the semantic value of conditional sentences. What she pushes is primarily the negative claim that conditionals do not have propositions as their semantic values. This is an important claim for the non-propositional view

²² Kratzer (1981) provides a battery of arguments that modals themselves need ordering (see Swanson, 2008, for critical review). Lewis (1981) proved the equivalence between the structure of Kratzer's semantics and his own.

since if conditionals did have propositions as their semantic values we would expect the normal bridging principles to kick in so that assertions of conditionals would simply be assertions of propositions.

For any given claim about what belief and assertion of conditional sentences amounts to, there will be a host of different combinations of semantic values and bridging principles that support that claim. So it is not obvious why you should choose one particular combination; if your main aim is to say what assertions of conditionals and belief in conditionals amount to, then it may be wise to remain neutral on which semantic values and bridging principles you think are correct. This is not to say that giving semantic values and bridging principles for conditionals is not an interesting project for those sympathetic to the non-propositional view, it is just to say that not everyone who argues for the non-propositional view needs to engage in it.

Nonetheless, if we are going to show that non-propositional views are compatible with the restrictor view we need to sketch how. This is what I turn to now with two different non-propositional semantics for conditionals, a trivalent account and a covert modal account.

3.10 Trivalent

Belnap (1970) gives a trivalent semantics for conditionals and a semantics for quantifiers that allows quantifiers to take trivalent formulas as their sole argument. The trivalent semantics is the usual one: $A \rightarrow C$ has the truth value of C when A is true and otherwise is undefined.²³ If there is an open variable, x , in $A \rightarrow C$ then we can quantify over conditionals with quantifiers defined like this:

- (12) **Most** _{x} φ is true iff for most objects o s.t. $\varphi x \rightarrow o$ is defined, $\varphi x \rightarrow o$ is true.

The technical point is that a trivalent conditional can encode both the restriction (i.e. where it is defined) and the truth values when the restriction is satisfied. So, it is possible to get a unary quantifier that takes a single trivalent formula that is equivalent to a restrictive binary quantifier that takes two bivalent formulas.

Since trivalent formulas do not correspond to ordinary propositions, they can act as a plausible *semantic value* of an indicative conditional for a non-propositional account.²⁴ There is no need to posit a covert modal operator for bare indicative conditionals, then. The trivalent semantic value still leaves open what personal-level account we give of conditionals; that depends on what bridging principles we use. The trivalent semantics is compatible, for instance, with Edgington's view of assertion of conditionals as suppositional/conditional assertion.²⁵

²³ There are a number of different options for what to do when A or C is undefined, but these aren't relevant here.

²⁴ Of course, this is a terminological issue: you might think trivalent truth-conditions do correspond to ordinary propositions. However, given the work they do here, that does not seem to be the right way to divide up the space of possibilities for conditionals.

²⁵ The crucial point is that the trivalent semantic value has enough information to both retrieve the supposition (the worlds where the semantic value is either true or false) and the division of the supposed worlds into those where the conditional is true and those where it is false.

Let me illustrate these points by going through a simple example. Take the sentence ‘If Mary is here, she is angry.’ On the trivalent view this has as its semantic value something that is true in worlds in which Mary is here and she is angry, false in which Mary is here and she is not angry, and undefined in worlds in which Mary is not here. Suppose we take as basic the notions of conditional assertion and conditional belief, as Edgington seems to. Then our bridging principles for assertions and belief can be stated as follows: if φ has a trivalent semantic value, then (1) an assertion of φ is a conditional assertion of the proposition that φ is true given that φ is defined and (2) a belief in φ is a conditional belief that φ is true, given that φ is defined.²⁶

3.11 Non-propositional Modals

The trivalent route is not the only non-propositional view of conditionals. Recent work on epistemic modals has resulted in a variety of proposals according to which sentences with epistemic modals do not express propositions (Yalcin, 2007; Swanson, 2006). We can combine these non-propositional views of modals with Kratzer’s hypothesis that bare conditionals contain silent necessity modals to get a non-propositional view of bare conditionals. This view needs three components:

- syntax/semantics of ‘if’ clauses are restrictors of modals;
- silent epistemic necessity modals in indicative conditionals like (10);
- non-propositional semantics for epistemic modals which can allow restrictions

To make the view complete we also need to posit bridging principles between the non-propositional values for epistemic modals and the personal-level notions relating to them such as assertion and belief. Yalcin (2007) and Swanson (2006) provide both of these in their compositional systems.

While Yalcin (2007) does not endorse the restrictor view, the semantic values he assigns to bare indicative conditionals and epistemic modals are available to someone with the restrictor view. (I give this variation on Yalcin’s semantics in the second appendix.)

3.12 Conditional Commands

Treating philosophical views of conditionals as theories of the meaning of entire sentences with bare conditionals can help clarify some issues about conditional commands. Edgington (2008) makes the following argument against the material conditional account of conditionals:

Conditional commands can [...] be construed as having the force of a command of the consequent, conditional upon the antecedent’s being true. The doctor says to the nurse in the emergency ward, ‘If the patient is still alive in the morning, change the dressing.’ Considered as

²⁶ It is worth noting that these bridging principles do not work well for other proposed instances of trivalence, such as that arising from vagueness: when I say that someone is tall, I do not assert that he is clearly tall, conditional on him not being a borderline case. An adequate trivalent semantics for conditionals and vagueness would need somehow to avoid this problem. This relates to the problems Soames (1989) raised for trivalent accounts of presupposition projection.

a command to make Hook's conditional true, this is equivalent to 'Make it the case that either the patient is not alive in the morning, or you change the dressing.' The nurse puts a pillow over the patient's face and kills her. On the truth-functional interpretation, the nurse can claim that he was carrying out the doctor's order. Extending Jackson's account to conditional commands, the doctor said 'Make it the case that either the patient is not alive in the morning, or you change the dressing', and indicated that she would still command this if she knew that the patient would be alive. This doesn't help. The nurse who kills the patient still carried out an order. Why should the nurse be concerned with what the doctor would command in a counterfactual situation?

Edgington is correct to find conditional commands puzzling if we think the material conditional account ['Hook'] is correct. However, even an advocate of the material conditional view of bare conditionals is entitled to a more sophisticated account of conditional commands if he endorses the restrictor view of 'if'. The obvious direction to go is to assume that imperatives include some sort of modal operator, and that the antecedent in a conditional command restricts this operator. If some account like this works, then the material conditional as a view about full sentences is completely compatible with an account of conditional commands that does not reduce them to material conditionals in the way Edgington suggests.

3.13 The Argument from Probability and Restrictors

So far, we have not seen any serious impact of the semantic insights of Lewis and Kratzer on the philosophical debate over conditionals, even on the debate between propositional and non-propositional views. In this section, I want to explore one way in which the restrictor view can be used to undermine an argument for the non-propositional view.²⁷

There is a well-known argument that goes from a simple observation about the probabilities that we assign to conditionals to the view that conditionals do not express propositions. The observation about the probabilities of conditionals is often called *Adams's Thesis*, the view that the probability of a conditional is its conditional probability, formally $P(A \rightarrow C) = P(C|A)$. Suppose we accept Adams's thesis. There are a number of simple mathematical results demonstrating that there is no proposition whose probability satisfies Adams's thesis. These results always depend on auxiliary assumptions of various sorts, but there is a wide-literature suggesting these assumptions are minimal and plausible.²⁸ So, the argument goes, $A \rightarrow C$ cannot be a proposition since there is no proposition that has the same probability as we think it does.

The restrictor view can undermine this argument for the non-propositional view by undermining some of the motivation for Adams's thesis. Recall that according to

²⁷ Some of the points here can be found in Cozic and Égré (2010) and Rothschild (2013), as well as in von Fintel's unpublished lectures (e.g. von Fintel, 2007).

²⁸ This literature begins with Lewis's (1976) famous triviality results; further stronger results are discussed in Edgington (1995) and, more formally, in Hájek and Hall (1994). Cozic and Égré (2010) make an important connection between the triviality results and the limitations of unary quantification referred to in note 12.

Adams's thesis the probability we assign to an indicative conditional is the probability of its consequent given its antecedent. One consideration in favor of Adams's Thesis goes by way of sentences like (13).

- (13) It's likely that if Mary is here, she is angry.

It seems (13) is something we would believe/assert just in case the probability that we assign to Mary being angry on the condition that she is here is high (see (5) above). How do we explain this fact? Well, Adams's thesis would explain it nicely: for on Adams's thesis whether or not we think an indicative conditional is likely just depends upon whether or not we think the consequent is likely given the antecedent. In this way Adams's thesis *explains* how we understand sentences like (13), and this itself is a consideration in favor of Adams's thesis.

The explanatory use of Adams's thesis above depends on the assumption that (13) involves an ascription of probability to an indicative conditional. The restrictor hypothesis, however, would favor a different account of the semantic structure of (13). On the restrictor hypothesis this is a classic instance in which an 'if'-clause restricts a probability operator. The probability judgment is simply a judgment of the probability of the consequent *restricted* to the worlds in which the antecedent is true. Assuming a reasonable semantics of probability operators such as 'likely' this will be true just in case the conditional probability is greater than 0.5 (see Yalcin, 2010, for a comprehensive discussion of the semantics of probability operators).

To make clear: the reason this strategy is compatible with the rejection of Adams's thesis is that on this strategy we do not concede that indicative conditionals themselves conform to Adams's thesis. The strategy works rather by denying that our apparent judgments of the probabilities of conditionals are really judgments of the probabilities of the propositions expressed by bare conditionals. On Kratzer's full view, for instance, indicative conditionals have silent necessity modals and express propositions.

So, given the restrictor view of 'if'-clauses, our judgments about sentences like (13) do not provide support for Adams's thesis. However, all cases of graded belief do not involve explicit probability operators. We can simply have a high degree of confidence in the indicative conditional 'If Mary is here, she is angry', without explicitly saying or thinking (13). Our confidence in a conditional seems to depend just on our conditional confidence in the consequent given the antecedent: this is another piece of evidence in favor of Adams's thesis. For the restrictor view to undermine this consideration, more assumptions about how 'if' operates need to be made than are standard in the restrictor literature. In particular, we need to allow that 'if'-clauses can act not just to restrict linguistically present modals but also can restrict aspects of *thoughts* involving probabilistic belief. This idea has not been much explored but it seems a promising approach to explain intuitions supporting Adams's thesis without actually endorsing Adams's thesis. Note, however, that if we follow this strategy, we seem to be already accepting one of the main tenets of the non-propositional view: belief in conditionals does not directly target a proposition. I am not going to argue here that we should reject Adams's thesis. I just want to suggest that a case can be made that Adams's thesis, taken as a thesis about bare indicative conditionals, is an illusion that can be explained away once we acknowledge that 'if'-clauses are restrictors.

3.14 Embedded Conditionals

Another area where semantic theory connects up with the philosophical debate over the meaning of conditionals is in the question of how conditionals embed under quantifiers. So far, we've discussed only one way in which conditionals can be embedded: under probability operators, modals, and adverbs of quantification. The restrictor story seems to provide a clear unified analysis of 'if' in these embeddings: the 'if'-clause serves to restrict the operator. Given that the restrictor view is compatible with either propositional or non-propositional accounts of bare conditionals, these cases do not provide evidence for or against the idea that bare conditionals express propositions.

There are, however, a variety of constructions in which 'if'-clauses are embedded in more complex constructions. It is commonly noted that many embeddings of conditionals in complex constructions do not seem interpretable. Sentence (14), as Gibbard (1981) notes, is not easily comprehensible.

- (14) If Kripke was there if Strawson was, then Anscombe was there.

I want to put aside the question of the significance of the fact that many instances of embedded conditionals like this are hard to understand.²⁹ There are, in any case, many examples of embedded conditionals which are perfectly easy to understand. Here are some instances:

Conditionals under conjunctionals:

- (15) If Mary is here then John is here, and John might be here.

Conditionals under disjunction:

- (16) Either if Mary is in China then she's in danger or if Mary is in India then she's in danger.

Conditionals under quantifiers (Higginbotham, 1986):

- (17) Some student will fail if he goofs off.

All of these sentences with embedded conditionals are easily comprehensible. I will focus on the cases of conditionals embedded under quantifiers, such as (17), as it is perhaps the best studied example.³⁰

Some, such as Kölbel (2000) argue that sentences of the form of (17) provide evidence against the non-propositional view. Kölbel argues that the problem embeddings of conditionals raise is analogous to the Frege–Geach problem for expressivism. That problem, generally speaking, is the problem of accounting for how

²⁹ Should we follow Gibbard (1981) and Edgington (1995, 2008) in seeing this as itself evidence for the non-propositional view? It is not clear to me that we should. After all, if the non-propositional views need to account for *some* embeddings, then they would seem also to face the problem of explaining the lack of generality. Of course, if they had a *predictive* theory about when exactly embeddings were acceptable, that could be an advantage, but I know of no such theory.

³⁰ Conjunctions, in any case, do not present serious problems for any accounts, given that conjunctions can be paraphrased as consecutive assertions. Disjunctions of conditionals would seem (from a logical point of view) to present similar issues to those raised by existential quantifiers.

sentences that do not express propositions function under the standard truth-functional operators (for a recent review see Schroeder, 2008).

Assuming that cases like (17) are genuine cases of conditionals embedded under operators, nothing prevents the non-propositional approaches from giving extended semantics for the relevant operators to try to cover these cases. The non-propositional approach assigns non-propositional semantic values to conditionals, so all that is needed is to expand the meaning of the quantifiers to allow embeddings of non-propositional values. Of course, doing so requires a number of theoretical choices, in particular the assignment of particular semantic values to conditionals. Swanson (2006) aims to give exactly such an account of examples like (17) as well as other embeddings. An important point here is that it is already standard practice in linguistics to allow basic logical operators to operate on a range of different types of semantic values, so that extending the meaning of the quantifiers and logical connectives is by no means unorthodox, if done in a principled and systematic way (Partee and Rooth, 1983; Partee, 1995).³¹

One theoretical option for treating quantified conditionals, available to propositional or non-propositional theorists who endorse the restrictor view, is to see 'if'-clauses as directly restricting nominal quantifiers. Supporting this view is the seeming equivalence of the following two sentences (as noted by Higginbotham, 1986):

- (18) a. Every student passed the exam if he tried.
- b. Every student who tried passed the exam.

This equivalence would be neatly explained by positing that 'if he tried' simply restricts the nominal quantificational phrase 'every student'. For then the logical form of (18-a) would be as in (19), which is clearly equivalent to (18-b).

- (19) Every [student & tried] [passed the exam]

This option has been explored recently (von Fintel, 1998; Leslie, 2009). However a systematic examination of cases suggests that we cannot hold that *generally* 'if'-clauses can restrict nominal quantifiers. If they could, we would expect (17) to have a reading on which it is equivalent to (20)

- (20) Some student who goofs off will fail.

It does not, however, which should make us suspicious of the idea that 'if'-clauses really can restrict nominal quantifiers such as 'every' and 'no'. For this and other reasons, the leading consensus is that accounting for the equivalence of (18-a) and (18-b) by appeal to the idea that 'if'-clauses restrict nominal quantifier is wrong (von Fintel and Iatridou, 2002; Huitink, 2010; Klinedinst, 2011).

³¹ In Yalcin's (2007) semantics for instance, the non-propositional nature comes in only through the interpretation of an index of evaluation. Thus, on his account we can simply use the off-the-shelf interpretation of all logical operators and get a complete semantic system. The interesting question is whether the semantic values we get when we do this, combined with the relevant bridging principles, provide a plausible account of the constructions. Klinedinst and Rothschild (2012) give cases where they do not and propose some fixes.

Since direct restriction is not an option, embedded conditionals under quantifiers provide serious challenges for any semantic account of conditionals. It is not sufficient to merely assign *some* semantic value to embedded conditionals. We also want the semantic value assigned to match our judgments about what the sentence means. For instance, the material conditional view allows us to assign propositions to the embedded conditionals in (15)–(17), but no matter how we construe the logical form of these sentences it does not seem like we will get the right truth-conditions for these sentences.³²

Nonetheless a serious effort has been made to show that the strict-conditional view (or a variably-strict view) gives adequate truth conditions for most instances of quantified conditionals. von Stechow and Iatridou (2002); Klinedinst (2011) show that a strict/variable strict conditional account can explain subtle facts about the meaning of quantified conditionals, such as the seeming equivalence between (21-a) and (21-b).

- (21) a. No student will pass if he goofs off.
- b. Every student will fail if he goofs off.

The basic idea is that the logical form of both sentences involves the embedding of a bare conditional in the matrix clause of the quantifier as follows:

- (22) a. No [student x] [if x goofs off, x will pass].
- b. Every [student x] [if x goofs off, x will fail].

If we now assume that the conditional excluded middle holds, i.e. in every case either $A \rightarrow C$ is true or $A \rightarrow \neg C$ is true, then the equivalence of (21-a) and (21-b) follows immediately. What is important to note is that this explanation of what is going on with the sentences such as (21-a) and (21-b) depends on conditionals expressing truth-valued propositions. At this point, then, propositional views would seem to have an advantage in treating quantified conditionals, but this is perhaps just a result of the fact that propositional theorists have worked more seriously on quantified conditionals than non-propositional theorists have.

Let me strengthen the consideration above by giving another case for which handling an embedded conditional is tractable on a propositional view but does not seem to be so on a non-propositional view. Consider sentence (23) in which a quantified conditional is embedded under probability operator.

- (23) It's likely that some student will pass if he tries.

Focus on the reading of (23) in which it means that there is a high chance that at least one student is such that were he to take the exam he would pass. How do we capture this reading in our semantics? We cannot view this sentence as one where the probability operator 'it's likely' is restricted by the 'if'-clause. For (23) is not equivalent to either of the two readings which we can get if we restrict the probability operator by the 'if'-clause (the two readings depend on the scope of 'some').

³² For this reason I share with Edgington (1995) perplexity over why the existence of embeddings is so often used to argue for the material conditional account

- (24) The conditional probability that a student will pass, given that some student takes the test is high.
- (25) There is some student x such the conditional probability that x will pass given that x takes the test is high.

(24) requires that there actually be a high chance that if any students take the test one student will pass, which is not the intended reading (for it might be unlikely that the one student who would pass were he to take the test will actually take it). On the other hand, (25) requires that we be certain that there is one student who will likely pass if he takes the test, which is also not the intended reading. It seems safe to say, then, that we cannot explain the natural reading of (23) by allowing the ‘if’ to restrict ‘it’s likely’.

A natural explanation of what is going on in (23) is as follows:

For every student x there is a proposition expressed by the sentence ‘if x tries, he will pass’. (23) is true just in case it is likely that one of those propositions is true.

If we accept this explanation, however, we are accepting that there is some *proposition* corresponding to the sentence ‘if x tries, he will pass’ for each x .³³ We need a proposition here because propositions are the sorts of thing we can assign probabilities to. If we accept that bare indicative conditionals (when embedded) can sometimes express propositions, then we have already rejected the non-propositional view in some cases.³⁴

3.15 Conclusion

My goal in this chapter was to relate the philosophical debate over conditionals to the linguistic literature on conditionals. In philosophy non-propositional views are both widely accepted and widely viewed with suspicion as being incompatible with the project of formal semantics. I argued here that we should not be so suspicious of non-propositional views, but I also suggested some challenges the views face.³⁵

Appendix 3.1: Restrictor Semantics

This is a simple syntactic variant of the restrictor view. It is meant to cover conditionals under adverbs of quantification, modals, and bare conditionals. We have two classes of expressions: sentences, which are true or false relative to situations (which can be actual or possible), and situational quantifiers (including modals), which are binary quantifiers taking a restrictor sentence and a matrix sentence.

³³ On the restrictor view we might get that proposition by restricting ‘will’ (or a silent necessity modal) in ‘if x takes the test he will pass’. The point I am making here is that the result of this process still yields a proposition which we can assign a probability to. This is exactly what the non-propositional view of bare conditionals seeks to deny.

³⁴ For one of the only attempts to deal with this general kind of example from a non-propositional perspective see Moss (2015).

³⁵ This note is intended as a supplement to, rather than a review of, the debate between propositional and non-propositional views, and so I have not discussed many crucial issues such as the alleged *subjectivity* of conditionals. I discuss this and related issues in Rothschild (2015).

First, the semantic rules for sentences:

- (26) $\llbracket \text{Mary is here} \rrbracket^s$ is true iff Mary is here in situation s .
 (27) $\llbracket \text{Mary is angry} \rrbracket^s$ is true iff Mary is angry in situation s .

Now, the semantic rules for the situational quantifiers:

- (28) $\llbracket \text{usually } [\varphi][\psi] \rrbracket^s$ is true iff in the world of s most situations, s' , in which $\llbracket [\varphi] \rrbracket^{s'}$ is true are situations in which $\llbracket [\psi] \rrbracket^{s'}$ is true.
 (29) $\llbracket \text{necessarily } [\varphi][\psi] \rrbracket^s$ is true iff for all epistemically-possible-in- s -situations, s' , if $\llbracket [\varphi] \rrbracket^{s'}$ is true then $\llbracket [\psi] \rrbracket^{s'}$ is true.

This is our basic semantics. Now we need to give our syntactic construal rules, which allow us to handle sentences that include ‘if’ (which is itself not interpreted). For these rules we need a special tautological sentence T . Let Q be one of the situational quantifiers. Let φ and ψ be sentences without ‘if’ appearing in them. The syntactic construal rules are as follows:

- (30) $Q\varphi \Rightarrow Q[T][\varphi]$.

This rule tells us that if a non-conditional sentence is embedded under an adverb of quantification, then it goes into the matrix of the adverb of quantification and the restrictor is vacuous, i.e. it is T .³⁶

- (31) $Q \text{ if } \varphi, \psi \Rightarrow Q[\varphi][\psi]$.

This rule tells us that if an adverb of quantification heads a conditional, then the antecedent becomes the restrictor and the consequent becomes the matrix.

- (32) $\text{if } \varphi, \psi \Rightarrow \text{necessarily if } \varphi, \psi$.

This rule adds a silent necessity modal to a bare conditional.

On these rules then sentence (33), by rules (31) and (32), is transformed into (34), which is true iff all epistemically-possible-in- s situations in which Mary is here are ones in which she is angry, which is a standard strict conditional.

- (33) $\llbracket \text{if Mary is here, she is angry} \rrbracket^s$.
 (34) $\llbracket \text{Necessarily [Mary is here][Mary is angry]} \rrbracket^s$.

Appendix 3.2: Non-propositional Modal Restrictor Semantics

This is a simple modification of the above semantics to yield Yalcin’s (2007) non-propositional view of conditionals.

Following Yalcin we add an extra index i , which is an information parameter, and has as its values sets of situations.³⁷ We consider sentences that express truth or falsity in a way that is sensitive to the information parameter to be *non-propositional* (see Yalcin, 2007, for discussion).

Syntactic construal rules are the same as before, as are all truth definitions, except that for *necessarily*. The new entry for *necessarily* (which is understood as an epistemic modal) is as follows:

³⁶ See von Stechow (1994, 1995) for discussion of pragmatic restrictions of adverbs of quantification.

³⁷ Yalcin uses worlds, not situations. I do not do this here to emphasize the parallels with adverbs of quantification.

- (35) $\llbracket \text{necessarily } [\varphi][\psi] \rrbracket^{s,i}$ if for all situations s' in i in which $\llbracket \varphi \rrbracket^{s',i}$ is true, $\llbracket \psi \rrbracket^{s',i}$ is true.

We now need a bridging principle for sentences whose truth is sensitive to i , such as sentences that include *necessarily*. The principle we will give, (36), takes sentences as recommendations to update one's belief states to make the sentence true if one's belief state is used to select i :

- (36) If an assertion of a sentence φ is made and φ is sensitive to i then that assertion should be understood as a recommendation to conform one's beliefs to the constraint: if b = set of all situations worlds possible according to one's belief state, then for all s in b , $\llbracket \varphi \rrbracket^{s,b}$ is true.

Consider:

- (37) $\llbracket \text{if Mary is here, she is angry} \rrbracket^{s,i}$.

By the syntactic construal rules this comes out as follows:

- (38) $\llbracket \text{Necessarily } [\text{Mary is here}][\text{Mary is angry}] \rrbracket^{s,i}$.

This is true iff all situations s in i in which Mary is here are situations in which Mary is angry. Since it is sensitive to the information parameter, it is non-propositional.³⁸

References

- Adams, E. W. (1975) *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: D. Reidel.
- Barwise, J. and Cooper, R. (1981) Generalized quantifiers and natural language. *Linguistics and Philosophy* 4: 159–219.
- Belnap, N. (1970) Conditional assertion and restricted quantification. *Noûs* 4: 1–12.
- Bennett, J. (2003) *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.
- Cozic, M. and Égré, P. (2010) If-clauses and probability operators. *Topoi* 30: 17–29.
- de Finetti, B. (1936) La logique de la probabilité. *Actes du congrès international de philosophie scientifique*, Fasc. IV: 31–9.
- Edgington, D. (1995) On conditionals. *Mind* 104(414): 235–329.
- Edgington, D. (2008) Conditionals. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Winter 2008 edition.
- Ellis, B. (1978) A unified theory of conditionals. *Journal of Philosophical Logic* 7: 107–24.
- Fintel, K. von (1994) *Restrictions on Quantifier Domains*. Ph.D. thesis, University of Massachusetts, Amherst.
- Fintel, K. von (1995) A minimal theory of averbial quantification. In B. Partee and H. Kamp (eds), *Context Dependence in the Analysis of Linguistic Meaning: Proceedings of the Workshop in Prague, February 1995, Bad Teinach, May 1996*, IMS Stuttgart Working Papers, pp. 153–93.
- Fintel, K. von (1997/2005) How to count situations. Unpublished note.
- Fintel, K. von (1998) Quantifiers and 'if'-clauses. *Philosophical Quarterly* 48: 209–14.
- Fintel, K. von (2007) If: The biggest little word. Slides from talk at Georgetown University Roundtable.
- Fintel, K. von and Iatridou, S. (2002) If and when *If*-clauses can restrict quantifiers. Paper for the Workshop in Philosophy and Linguistics at the University of Michigan.

³⁸ Note this semantics is only meant to deal with a fragment of the language; it does not give an adequate treatment of embeddings of non-propositional material in antecedents, for instance.

- Geis, M. L. (1970) *Adverbial Subordinate Clauses in English*. Ph.D. thesis, MIT.
- Gibbard, A. (1981) Two recent theories of conditionals. In W. L. Harper, R. Stalnaker, and G. Pearce (eds), *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: Reidel.
- Gillies, A. (2010) Iffiness. *Semantics and Pragmatics* 3: 1–42.
- Grice, P. (1967/1989) Logic and conversation. In *Studies in the Ways of Words*. Cambridge, MA: Harvard University Press.
- Groenendijk, J. and Stokhof, M. (1984) *Studies in the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, University of Amsterdam.
- Hájek, A. and Hall, N. (1994) The hypothesis of the conditional construal of conditional probability. In E. Eells, B. Skyrms, and E. Wilcox Adams (eds), *Probability and Conditionals: Belief Revision and Rational Decision*. Cambridge: Cambridge University Press.
- Hamblin, C. (1973) Questions in Montague English. *Foundations of Language* 10: 41–53.
- Heim, I. (1982) *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts, Amherst.
- Higginbotham, J. (1986) Linguistic theory and Davidson's program in semantics. In E. LePore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Oxford: Blackwell.
- Huitink, J. (2008) *Modals, Conditionals and Compositionality*. Ph.D. thesis, Radboud Universiteit Nijmegen.
- Huitink, J. (2010) Quantified conditionals and compositionality. *Language and Linguistics Compass* 4: 42–53.
- Jackson, F. (1987) *Conditionals*. Oxford: Blackwell.
- Karttunen, L. (1977) Syntax and semantics of questions. *Linguistics and Philosophy* 1: 3–44.
- Khoo, J. (2011) Operators or restrictors? A reply to Gillies. *Semantics and Pragmatics* 4(4): 1–43.
- Klinedinst, N. (2011) Quantified conditionals and conditional excluded middle. *Journal of Semantics* 28: 149–70.
- Klinedinst, N. and Rothschild, D. (2012) Connectives without truth-tables. *Natural Language Semantics* 20: 137–75.
- Kölbel, M. (2000) Edgington on compounds of conditionals. *Mind* 109: 97–108.
- Kratzer, A. (1978) *Semantik der Rede. Kontexttheorie, Modalwörter, Konditionalsätze*. Königstein: Scriptor.
- Kratzer, A. (1981) The notional category of modality. In H.-J. Eikmeyer and H. Reiser (eds), *Words, Worlds, and Contexts*. Berlin: Walter de Gruyter, pp. 38–74.
- Kratzer, A. (1986) Conditionals. *Chicago Linguistics Society* 22(2): 1–15.
- Kratzer, A. (2011) Situations in natural language semantics. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Fall 2011 edition.
- Kratzer, A. (2012) *Modals and Conditionals*. Oxford: Oxford University Press.
- Leslie, S.-J. (2009) 'If', 'unless', and quantification. *Studies in Linguistics and Philosophy* 85: 3–30.
- Lewis, D. (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lewis, D. (1975) Adverbs of quantification. In E. L. Keenan (ed.), *Formal Semantics of Natural Language*. Cambridge: Cambridge University Press.
- Lewis, D. (1976) Probabilities of conditional and conditional probabilities. *Philosophical Review* 8: 297–315.
- Lewis, D. (1980) Index, context, and content. In S. Kranger and S. Ohman (eds), *Philosophy and Grammar*. Dordrecht: Reidel, pp. 79–100.
- Lewis, D. (1981) Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic* 10: 217–34.
- Lewis, D. (1986) Probabilities of conditional and conditional probabilities II. *The Philosophical Review* 95: 581–9.

- Moss, S. (2015) The semantics and pragmatics of epistemic vocabulary. *Semantics and Pragmatics* 8(5): 1–81.
- Ninan, D. (2010) Semantics and the objects of assertion. *Linguistics and Philosophy* 33(5): 355–80.
- Partee, B. (1995) Many quantifiers. In E. Bach, E. Jelinek, A. Kratzer, and B. Partee (eds), *Quantification in Natural Languages*. London: Kluwer.
- Partee, B. H. and Rooth, M. (1983) Generalized conjunction and type ambiguity. In E. Bäuerle (ed.), *Meaning, Use, and Interpretation of Language*. Berlin: de Gruyter.
- Rabern, B. (2012) Against the identification of assertoric content with compositional value. *Synthese* 189(1): 75–96.
- Rothschild, D. (2013) Do indicative conditionals express propositions? *Nous* 47: 49–68.
- Rothschild, D. (2014) Capturing the relationship between conditionals and conditional probability with a trivalent semantics. *Journal of Applied Non-Classical Logics* 24: 144–52.
- Rothschild, D. (2015) Conditionals and propositions in semantics. *Journal of Philosophical Logic* 44(6): 781–91.
- Schroeder, M. (2008) What is the Frege–Geach problem? *Philosophy Compass* 3: 703–20.
- Soames, S. (1989) Presuppositions. In D. Gabbay and F. Guenther (eds), *Handbook of Philosophical Logic*, vol. IV, Dordrecht: D. Reidel, pp. 553–616.
- Stalnaker, R. (1968) A theory of conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, Oxford: Blackwell, pp. 98–112.
- Stanley, J. (2002) Modality and what is said. *Philosophical Perspectives* 16: 321–44.
- Swanson, E. (2006) *Interactions with Context*. Ph.D. thesis, MIT.
- Swanson, E. (2008) Modality in language. *Philosophy Compass* 3: 1193–1207.
- Yalcin, S. (2007) Epistemic modals. *Mind* 116: 983–1026.
- Yalcin, S. (2010) Probability operators. *Philosophy Compass* 5(11): 916–37.

4

Chasing Hook: Quantified Indicative Conditionals

Angelika Kratzer

I should say this upfront. The Hook from Edgington's *Conditionals* is a man with opinions. He thinks that *if* is a truth-functional connective and corresponds to material implication. My Hook is not a 'he' or a 'she', but an 'it'. It is material implication itself. It is $\supset$. Hook is elusive. We know it has a connection with *if*, but we don't quite know what the connection is. My project is to hunt Hook down in the back alleys of English. It's not that I think Hook is that special. I am interested in Hook because it makes a good probe for exploring the properties of embedded conditionals.

Embedded conditionals are a potential problem for theories of conditionals that don't give them truth-conditions. Lewis (1976) voiced an admittedly inconclusive objection: 'We think we know how the truth conditions of compound sentences of various kinds are determined by the truth conditions of constituent subsentences, but [if conditionals had no truth conditions A. K.] this knowledge would be useless when it comes to conditional subsentences.' Edgington (1995) accepts the first part of Lewis's statement, but then goes on: 'But this knowledge *is* useless when it comes to conditional subsentences. We do not have a satisfactory general account of sentences with conditional constituents' (p. 281). What's on the docket, then, are the prospects for a satisfactory general account of sentences with embedded conditionals.

4.1 Higginbotham's Puzzle

The embarrassment had been known for a long time, but nobody dared talk about it. Then Higginbotham (1986) dragged it into the open. Then many tried their hand at it (von Fintel, 1998; Dekker, 2001; von Fintel and Iatridou, 2002; Higginbotham,

* I am very indebted to Daniel Rothschild and Lee Walters for extensive comments on an earlier version of this chapter. The research leading to the chapter has benefitted from funding and intellectual stimulation provided by the University of Massachusetts at Amherst, the Radcliffe Institute for Advanced Study at Harvard University, and (via François Recanatì) by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013)/ERC grant agreement n° 229 441 -CCC. I am also grateful to audiences at EHESS and the Institut Jean Nicod in Paris (2009), Rutgers University (Ernie Fest, 2010), LACSI (NCSU, 2011), the Chicago Linguistics Society (2011), the Central APA (Chicago, 2012), and the University of Konstanz (*What if* collaborative research group, 2013).

2003; Abbott, 2004; Égré and Cozic, 2008; Leslie, 2009; Huitink, 2009; Klinedinst, 2011). The embarrassment was for those who believe in compositionality. Examples 1(a) and (b) are (almost) Higginbotham's.

- (1) a. Everyone will fail if they goof off.
- b. No one will pass if they goof off.

If failing is not passing, 1(a) and (b) are equivalent. They express the same proposition. 1(a) and (b) are also syntactic isomorphs; they are put together in the same way. And their semantic type trees are the same. The meanings of 1(a) and (b) should be put together in the same way, too, then. But that can't be so if *if* means Hook. Hook does well with 1(a), but fails badly with (b). 2(b) comes out true iff everyone goofs off and no one succeeds. But 1(b) does not require that everyone goof off, and so 2(b) cannot be the correct formalization of 1(b).

- (2) a. $\forall x (\text{goof-off}(x) \supset \text{fail}(x))$ Correct formalization of 1(a).
- b. $\neg \exists x (\text{goof-off}(x) \supset \text{pass}(x))$ Incorrect formalization of 1(b).

Some say there is a silent *always* that comes with *if* in cases like 1(a) or (b), or a silent necessity modal. That's no help. 3(a) and (b) still aren't equivalent, nor are 4(a) and (b).

- (3) a. $\forall x \forall t (\text{goof-off}(x)(t) \supset \text{fail}(x)(t))$
- b. $\neg \exists x \forall t (\text{goof-off}(x)(t) \supset \text{pass}(x)(t))$
- (4) a. $\forall x \forall w (\text{goof-off}(x)(w) \supset \text{pass}(x)(w))$
- b. $\neg \exists x \forall w (\text{goof-off}(x)(w) \supset \text{succeed}(x)(w))$

Some say that *if*-clauses restrict quantificational operators. If they can restrict adverbial or modal quantifiers, why not determiner quantifiers, too? The *if*-clauses in 1(a) and (b) should then restrict the domains of *every* and *no* in the same way a restrictive relative clause would. 1(a) should mean the same as 5(a), and 5(b) should be a paraphrase of 1(b).

- (5) a. Everyone who goofs off will fail.
- b. No one who goofs off will pass.

Edgington (2000), responding to Kölbel (2000), observed that sentences with embedded conditionals do not generally have equivalent paraphrases with restrictive relative clauses. Higginbotham (2003), Leslie (2009), and von Stechow and Iatridou (2002) made the same point.

Leslie imagines a student, Meadow, whose teacher would never fail her, regardless of how well she did. Meadow happens to work very hard for that teacher's class, however, and is thus not among those who goof off. The relative clause *who goofs off* in 5(a) and (b) can be readily understood as restricting the domain of *everyone* and *no one* to those students who are actually in the habit of goofing off or will actually goof off in the course of the class. Since Meadow is not among them, she is no obstacle to the truth of 5(a) or (b). In contrast, the *if*-clause in 1(a) and (b) creates a strong pull towards an interpretation where we look at all students in turn and consider situations where they goof off, moving on to merely possible situations if there aren't any actual ones. On that interpretation, the fact that Meadow would pass if she goofed off makes her a falsifying instance for 1(a) and (b).

von Fintel and Iatridou (2002) observe that the construction exemplified by 1(a) and (b) is constrained in a way that would be unexpected if *if*-clauses restricted determiner quantifiers in the way relative clauses do. We would have no obvious explanation for the contrast between 6(a) and 6(b) (modeled after Goodman, 1955), for example.

- (6) a. Every coin that is in my pocket is silver.
b. #Every coin is silver if it is in my pocket.

Unlike 6(a), 6(b) is odd. It suggests a non-accidental link between coins that are silver and coins in my pocket. The suggestion of a non-accidental link between antecedent and consequent is a well-known property of conditionals. It points to a complete conditional construction embedded under *every coin*.

Another striking contrast discovered by von Fintel and Iatridou is illustrated by the minimal pair 7(a) and (b) (their examples 27(a) and (b)).

- (7) a. Nine of the students will succeed if they work hard.
b. Nine of the students who work hard will succeed.

7(b) presupposes that there are more than nine students who work hard. 7(a) has no such presupposition. 7(a) says that there are nine students who 'have it in them' to succeed if they work hard. Nothing is implied about the number of students who actually work hard. This interpretation points again to a complete conditional construction in 7(a).

To sum up, we have seen evidence suggesting that sentences like 1(a) and (b) embed complete conditional constructions. They are genuine cases of embedded conditionals, then. In light of Edgington's (2001) verdict that 'no general algorithmic approach to complex statements with conditional components has yet met with success', the prospects for an insightful analysis of such constructions look daunting. The following two sections will feed Edgington's skepticism: The discussion will get to a point where the prospects for a general account of conditionals embedded under quantifier phrases look outright hopeless. The rest of the chapter will then begin to gather support for a more positive outlook.

4.2 Abbott's Puzzle

Assuming that passing is not failing, the logical make-up of 1(a) and (b) can be displayed as in 8(a) and (b):

- (8) a. $\forall x$ (if goof-off(x), fail(x))
b. $\forall x \neg$ (if goof-off(x), $\neg$ fail(x))

We can now see clearly that, to derive the equivalence of 1(a) and (b), we need a conditional that makes 9(a) and (b) equivalent.

- (9) a. (if goof-off(x), fail(x))
b. $\neg$ (if goof-off(x), $\neg$ fail(x))

Assuming a bivalent background logic, 9(a) and (b) are equivalent just in case 10(b) is the negation of 10(a).

- (10) a. (if goof-off(x), fail(x))
b. (if goof-off(x), $\neg$ fail(x))

The conditional we are looking for, then, needs to be one that is negated by negating its consequent. Like material implication, it doesn't allow opposite conditionals (like 10(a) and (b)) to be both false. It has to obey Conditional Excluded Middle (CEM). But unlike material implication, our conditional also doesn't allow opposite conditionals to be both true. It has to obey Weak Boethius' Thesis (WBT).¹

(11) $(\text{if } A, B) \vee (\text{if } A, \neg B)$ CEM

(12) $(\text{if } A, \neg B) \supset \neg(\text{if } A, B)$ WBT

On a material implication interpretation, the opposite conditionals 10(a) and (b) are both true if *x* doesn't actually goof off. Leslie's example shows that we are looking for a conditional that cares about what would happen if *x* were to goof off. As Higginbotham (2003) and Klinedinst (2011) point out, there is a conditional in the literature that almost fits the bill—Stalnaker's (Stalnaker 1968).

Stalnaker's conditional is true in a world *w* just in case its consequent is true in the closest world to *w* where its antecedent is true. If the antecedent is impossible, the closest world to *w* is stipulated to be the absurd world, where everything is true. On Stalnaker's analysis, conditionals with impossible antecedents are true, hence opposite conditionals with impossible antecedents are both true, violating WBT. The violation seems minor, though, and comes from a stipulation that feels a little arbitrary. We might set aside the impossible case. Stalnaker's conditional would then explain the intuitive equivalence of 1(a) and (b). Leslie's student Meadow is no longer a counterexample. Since she isn't actually goofing off we have to consider the closest world where she is. In that world she will still pass, hence 1(a) and (b) both wind up false. Stalnaker's analysis does well with conditionals like 1(a) and (b). But there are other kinds of conditionals that insist that we stick to actuality. The point was made in Abbott (2004). My example is a variation on one of hers, but is also different in important respects. It was constructed so as to not merely challenge the semantic side of Stalnaker's account of indicative conditionals. It also tries to block the possibility of invoking the pragmatic side of his account.²

Email Handling

You have two employees who, between them, are required to jointly answer all of your email queries, as long as they come from a respectable address. Your clients come from India or the US. One of the employees, Good Employee, handles all mail from India. She reliably answers all queries that come from respectable addresses. The other employee, Bad Employee, handles all mail from the US. She never answers any queries at all. There have been complaints recently about unanswered queries. You pick a particular time window for investigation: last month. Here is what you found. Good Employee handled 100 queries during that time. By sheer accident, they were all sent from respectable addresses, and she answered all of them. Bad Employee also handled 100 queries last month. By sheer accident, not a single one was sent from a respectable address, and she didn't answer any of them.

In the situation described, (13) is intuitively true.

(13) Every query was answered if it was sent from a respectable address.

¹ I am using the terminology of Pizzi and Williamson (2005).

² See Stalnaker (2014), ch. 6.

Assume that your clients all have both respectable and dubious e-mail addresses that they can access with equal ease, and that they don't send messages from any other country but their own. Assume whatever else it may take to make it so that every query that actually came from a dubious (not respectable) address would still have landed in Bad Employee's mailbox if it had come from a respectable address. (13) would then seem to wind up false on Stalnaker's account. We would seem to have to consider for every query that actually came from a dubious address the closest world where it didn't. The query would still land in Bad Employee's mailbox in that world, and thus remain unanswered.

But Stalnaker's account of indicative conditionals also has a pragmatic side to it. Interestingly, embedding a conditional under a quantifier phrase affects its pragmatic properties in crucial ways. For example, (13) can be completely acceptable and natural in contexts, where for each of the 200 queries under investigation, it is common knowledge whether or not it came from a respectable address. Here is an illustration. Suppose you assembled a list of the 200 queries, each paired with information about whether or not it came from a respectable address, and whether or not it was answered. You met with your two employees and put the list on the table. You could then use (13) as a way of summarizing the data you are looking at together. In this situation, the content of the list has become Common Ground among the three of you. There is no world in the Common Ground where a query that came from a dubious address in the actual world came from a respectable address. With respect to the assumed Common Ground, then, it was impossible for any of those dubious queries to come from a respectable address.

According to Stalnaker (1968), the truth of a conditional depends on a selection function f , which maps a proposition and a world to a selected world. A conditional of the form (*if A, B*) is true in a world w with respect to a selection function f just in case B is true in $f(A, w)$. Not just any selection function is permitted. There are constraints. One says that A must be true in $f(A, w)$. Another one lets the selection function pick the absurd world only if A is impossible. This condition is met on our scenario. In the context described above, the antecedent of 'if x was sent from a respectable address, x was answered' is impossible for any x that was actually sent from a dubious address. Once the pragmatic part of Stalnaker's analysis is taken into account, then, (13) comes out true in contexts like the one we imagined. That's right for our example, but it is not the end of the story.

In the assumed context, the embedded conditional of (13) is precisely the marginal case of a Stalnaker conditional that we had to set aside in our search for a conditional that makes 1(a) and (b) equivalent. Stalnaker's conditional only delivers the equivalence of 1(a) and (b) as long as we don't have to worry about impossible antecedents. Otherwise, WBT fails. The problem that we are now facing is that sentences like 14(a) and (b) below feel no less equivalent than 1(a) and (b), but we can no longer afford to neglect cases with impossible antecedents. The scenario we have been considering confronts us with just such a case.

- (14) a. Every query was answered if it was sent from a respectable address.
- b. No query was not answered if it was sent from a respectable address.

If material implication is ruled out as a possible interpretation for the embedded conditional in 14(a) and (b), so is the Stalnaker conditional. We are back to square one. We have no general analysis of conditionals embedded under quantifier phrases. Stalnaker's conditional delivers the equivalence of 1(a) and (b) under plausible assumptions, but those assumptions are no longer plausible for 14(a) and (b): the cases that can be neglected for 1(a) and (b) can no longer be neglected for 14(a) and (b).

4.3 Pizzi and Williamson's Bombshell

Let us take stock of where things stand. We convinced ourselves that for 14(a) and (b) to come out equivalent, the embedded conditional must obey both CEM and WBT.

- (14) a. Every query was answered if it was sent from a respectable address.
 b. No query was not answered if it was sent from a respectable address.

We have also seen that the embedded conditional in 14(a) and (b) has different properties from that in 1(a) and (b). The *if*-clause in 1(a) and (b), but not that in 14(a) and (b), allows us to consider mere potentials. Tracking the behavior of the kind of conditional exemplified in 14(a) and (b) in other syntactic environments reveals moreover that, as a type, it validates both Modus Ponens (MP) and Contraposition (CP).

- (15) (If A, B), A $\vdash$ B MP
 (16) (If A, B) $\vdash$ (if $\neg$ B, $\neg$ A) CP

Modus Ponens used to be considered a solid principle for reasoning with conditionals, but then there was McGee (1985), who came with a series of counterexamples. Here is the best-known one (McGee, 1985): 462).

McGee

Opinion polls taken just before the 1980 election showed the Republican Ronald Reagan decisively ahead of the Democrat Jimmy Carter, with the other Republican in the race, John Anderson, a distant third. Those apprised of the poll results believed, with good reason:

If a Republican wins the election, then if it's not Reagan who wins it will be Anderson. A Republican will win the election.

Yet they did not have reason to believe that if it's not Reagan who wins, it will be Anderson.

The conditionals discussed in McGee (1985) either contain the modal *will* or have the feel of lawlike generalizations. They very readily comply with Modus Ponens as soon as we change them into the bare past tense conditionals exemplified in (14).

Transposed McGee

Opinion polls taken just before the 1980 election showed the Republican Ronald Reagan decisively ahead of the Democrat Jimmy Carter, with the other Republican in

the race, John Anderson, a distant third. Imagine that the polls just closed and those apprised of the poll results believe, with good reason:

If a Republican won the election, then if it wasn't Reagan who won, it was Anderson. A Republican won the election.

They had every reason to believe that if it wasn't Reagan who won, it was Anderson.

Contraposition is usually rejected for counterfactuals, but seems to hold for the kind of indicative conditional we have in (14). If any breed of conditionals obeys Contraposition, that particular breed does.

- (17) a. If this query was sent from a respectable address, it was answered.
- b. If this query was not answered, it was not sent from a respectable address.

Yet sometimes, bare past conditionals like 18(a) and (b) are offered as potential counterexamples to Contraposition.

- (18) a. If Mary read the paper, she didn't read it this morning.
- b. If Mary read the paper this morning, she didn't read it.

Both 18(a) and (b) convey that Mary didn't read the paper this morning. 18(a) is acceptable, 18(b) is not. The difference doesn't seem to be a difference in truth-conditions, though. 18(a) is only appropriate in contexts where Mary might have read the paper (Stalnaker, 1975). Likewise, 18(b) should only be appropriate in contexts where Mary might have read the paper this morning. However, no cooperative speaker uttering 18(b) can possibly believe that Mary might have read the paper this morning, since 18(b) implies that she didn't read the paper this morning. 18(a) and (b) are an odd pair of contraposed conditionals, but they are no counterexample to Contraposition. I conclude that both Modus Ponens and Contraposition are valid for the type of indicative conditional exemplified in (14).

There is a bombshell hidden in the conclusion I just drew: it turns out that any conditional that satisfies MP, CP, CEM, and WBT is equivalent to the material biconditional. In other words, if the conditional embedded in 14(a) and (b) is a type of conditional that validates MP, CP, CEM, and WBT, we can prove that it has to be equivalent to the material biconditional. The proof is in Pizzi and Williamson (2005). It is easy to see that the material biconditional satisfies MP, CP, CEM, and WBT. The other direction requires a little more work. Following the strategy in Pizzi and Williamson, it can be shown that, assuming a bivalent background logic, 19(a) to (d) are valid for any conditional satisfying MP, CP, CEM, and WBT.

- (19) a. $(A \leftrightarrow B) \supset (\text{if } A, B)$
- b. $(A \leftrightarrow \neg B) \supset \neg(\text{if } A, B)$
- c. $(\neg A \leftrightarrow B) \supset \neg(\text{if } A, B)$
- d. $(\neg A \leftrightarrow \neg B) \supset (\text{if } A, B)$

If 19(a) to (d) are valid for a conditional, it is true just in case its antecedent and consequent are both true or both false. Those are the truth-conditions for the material biconditional.

What is upsetting about the result we have just derived is that the conditional embedded in 14(a) or (b) doesn't look or feel like a biconditional.³ Intuitively, 14(a) and (b) could be true in cases where some queries from dubious, not respectable, addresses were also answered. But that can't seem to be so if the embedded conditional is a material biconditional. I have slipped into what looks like a paradoxical situation. Where did I go wrong? I stand by every single step in the reasoning I went through. 14(a) and (b) do feel equivalent. The embedded conditionals should therefore satisfy CEM and WBT. Those particular kinds of indicative conditionals also satisfy MP and CP. Pizzi and Williamson's proof is correct. The embedded conditional in 14(a) and (b) has to be equivalent to the material biconditional, then. How can this be?

4.4 Solving Abbott's Puzzle: Hook in Hiding

Let us take stock again. We have zoomed-in on a breed of conditional that is different from *will*-conditionals. It validates Modus Ponens and Contraposition. When it is embedded under a quantifier phrase, it doesn't allow us to consider individuals' mere potentials for satisfying the antecedent. It looks just like Hook. The problem is that we also judge sentences like 14(a) and (b) as equivalent.

- (14) a. Every query was answered if it was sent from a respectable address.
- b. No query was not answered if it was sent from a respectable address.

Pizzi and Williamson's proof seems to establish that for 14(a) and (b) to be equivalent, the embedded conditional has to be the material biconditional. Our dilemma is that we do not perceive a material biconditional in 14(a) or (b).

There must be some element in the syntactic environment of the embedded conditionals in 14(a) and (b) that obscures their compositional meaning contribution. The meaning of 14(a) and (b) can't just be composed from a conditional and a quantifier phrase with a fixed, context independent, denotation. There has to be another player. I want to suggest that that other player is no stranger: it is whatever device is responsible for covert domain restrictions for nominal quantifiers. Many authors, including von Stechow (1994), Stanley and Szabo (2000), Stanley (2000), and Martí (2003), have posited domain variables to account for covert quantifier domain restrictions. Martí and von Stechow have argued moreover that nominal domain restriction variables are attached to determiners.

- (20) No_D query was not answered if it came from a respectable address.

The domain restriction variable in (20) needs a value, and the embedded *if*-clause is a natural provider. In out-of-the-blue contexts, it is the only possible provider. On this proposal, the *if*-clause in 14(a) and (b) plays a double role. It is the antecedent of an

³ Conditionals can sometimes be strengthened to biconditionals via a pragmatic process called 'Conditional Perfection' in the linguistic literature (Geis and Zwicky, 1971), see also the discussion of examples (25) to (27) below). Conditional Perfection might affect 14(a): the effect can be contextually manipulated, though, and doesn't affect 14(b) in the same way.

embedded conditional, while simultaneously restricting the domain of the nominal quantifier. Logical forms for 14(a) and (b) would amount to 21(a) and (b).

- (21) a. $(\forall x: \text{query}(x) \& \text{from-respectable-address}(x))$
 $(\text{from-respectable-address}(x) \supset \text{answered}(x))$
 b. $(\forall x: \text{query}(x) \& \text{from-respectable-address}(x))$
 $\neg(\text{from-respectable-address}(x) \supset \neg \text{answered}(x))$

21(a) and (b) have the right interpretation. They do not imply that no queries from dubious, not respectable, addresses were answered. They are equivalent, and boil down to (22).

- (22) $(\forall x: \text{query}(x) \& \text{from-respectable-address}(x)) \text{ answered}(x)$

The proposal preserves von Fintel and Iatridou's insight that a complete conditional is embedded under the quantifier phrase in constructions like 14(a) and (b). In this particular case, the embedded conditional seems to be Hook.

If *if*-clauses can only restrict nominal domains indirectly through the mediation of a domain variable that comes with DPs (determiner phrases), we might expect nominal domain restriction via *if*-clauses to be sensitive to the nature of those DPs. Probing into this question, we find that different types of DPs do indeed differ in their ability to be restricted by *if*-clauses. (23) and (24) illustrate.

- (23) a. Every query was answered if it came from a respectable address.
 b. No query remained unanswered if it came from a respectable address.
 c. Most queries were answered if they came from a respectable address.
 (24) a. #Exactly fifty queries were answered if they came from a respectable address.
 b. #At least fifty queries were answered if they came from a respectable address.
 c. #At most fifty queries were answered if they came from a respectable address.

In 23(a) to (c), the *if*-clauses can restrict the domains of the quantifiers with relative ease. In contrast, readings where the *if*-clauses restrict the domains of the quantifiers are hard to get, if not unavailable, for 24(a) to (c). As a result, we can't seem to figure out what those sentences say. We are confused about how to count the queries. Clear cases of verifying instances are those queries that came from respectable addresses and were answered. But what are we supposed to do with queries that came from dubious addresses? My mind wants to side with Nicod (1924) and veto them as confirming instances of 24(a) to (c). But it also seems to want to interpret the embedded conditional as Hook. It is caught in a conundrum, then, that it can't seem to resolve.

(23) and (24) sort DPs in familiar ways. According to Landman (2004), those in (24) are born with property interpretations ('denotations at the type of sets') that may be type-shifted into other denotations in particular syntactic environments. The property interpretation emerges in constructions like *the exactly fifty queries*, *the at*

most fifty queries, or *the at least fifty queries*, for example. In contrast, the DPs in (23) begin life with contentful determiners that map properties to generalized quantifiers. If nominal domain variables require contentful determiners to attach to, we have an explanation for the pattern in (23) and (24). More work is needed, of course, to put the assumption that nominal domain variables have to attach to contentful determiners on a more solid footing. I will have to leave that project for another occasion.⁴

A potential argument supporting the assumption that *if*-clauses restrict determiner quantifiers pragmatically, rather than semantically or syntactically, can be constructed by showing that discourse properties can be manipulated to create configurations where the consequents, rather than the antecedents, of conditionals act as restrictors for quantifier domains. (25) illustrates.

(25) You: Did you see kids using calculators when you volunteered in your son's school yesterday? What did they use the calculators for?

Me: *Most kids asked for calculators if they had to do long divisions*. But I am pleased to report that most kids in my son's school do long divisions by hand.

The targeted sentence in (25) is (26):

(26) Most kids asked for calculators if they had to do long divisions.

The context for (26) in (25) is set up so that the consequent of the embedded conditional is old information and the antecedent is new information. This manipulation has the effect of restricting the domain of *most* in (26) within (25) to kids who asked for calculators. The claim is that most kids who asked for calculators had to do long divisions. Only this interpretation of (26) is consistent with the sentence following it in (25), which adds the information that most kids in my son's school who do long divisions do not use calculators. Since my reply in (25) has an interpretation that feels entirely consistent, there must be an interpretation of (26) where the consequent, rather than the antecedent, of the embedded conditional restricts the domain of the embedding quantifier phrase.

The context manipulation that allows the consequent of the embedded conditional in (26) to be a restrictor has a second, well-known, effect. The emphasis placed on the antecedent creates an *only*-implicature, de facto turning the embedded conditional into a biconditional. The effect can be tracked more clearly with the unembedded conditional in (27):

(27) You: Does your son use a calculator in math classes? And if so, what does he use the calculator for?

Me: My son uses a calculator if he has to do long divisions.

In the context of (27), my reply is most readily understood as saying that my son uses a calculator if he has to do long divisions, but not otherwise. The suggestion is that if

⁴ An equally plausible story can be told if nominal quantifier domain restrictions are accounted for by routine situation arguments, rather than by special domain restriction variables. Here, there are already arguments that strong, but not weak, quantificational determiners introduce situation arguments. See Keshet (2008), Schwarz (2009, 2012), and Elbourne (2013) for discussion of this issue.

he doesn't have to do long divisions, he doesn't use a calculator. The pragmatic process that turns a conditional into a biconditional (Conditional Perfection (Geis and Zwicky, 1971)) has generated a huge literature and cannot be done justice here. For our current argument, it is important that there is such a process, that it is facilitated by placing emphasis on the antecedent, and that it can apply to embedded conditionals. A logical form that displays the intended interpretation of (26) is (28), which is equivalent to (28'):

(28) (Most x : $\text{kid}(x) \ \& \ \text{asked-for-calculator}(x)$) (has-to-do-long-divisions (x) $\equiv$ asked – for – calculator(x))

(28') (Most x : $\text{kid}(x) \ \& \ \text{asked-for-calculator}(x)$) has-to-do-long-divisions(x)

The idea that *if*-clauses may play both a semantic and a pragmatic role is neither new nor outlandish. von Fintel (2001) uses an example from Edgington (1995: 252f) to illustrate the pragmatic, dynamic, effect of *if*-clauses.

The Missing Hard Hat

For example, a piece of masonry falls from the cornice of a building, narrowly missing a worker. The foreman says: 'If you had been standing a foot to the left, you would have been killed; but if you had (also) been wearing your hard hat, you would have been alright.'

As Edgington notes (for a related example, p. 253; see also Frank, 1996), the order of presentation of the counterfactuals matters in such discourses. The first counterfactual in the previous passage feels true, but that very same counterfactual in the following passage comes across as false.

Transposed Missing Hard Hat

For example, a piece of masonry falls from the cornice of a building, narrowly missing a worker. The foreman says: 'If you had been wearing your hard hat, you would have been alright; but if you had been standing a foot to the left, you would have been killed.'

In both versions of the Missing Hard Hat, the antecedent of the first counterfactual remains active in the discourse and has a continued pragmatic effect beyond its semantic contribution to the truth-conditions of the first counterfactual. In the original Missing Hard Hat example, that effect can be channeled back into the semantics via the anaphoric particle *also*. In the transposed Missing Hard Hat example, the impact of the first antecedent is more indirect. It might update the set of worlds that are relevant for the interpretation of the second counterfactual, as on the accounts of von Fintel (1999, 2001) and Gillies (2007). Alternatively, it could make salient the possibility that the worker might have been wearing a hard hat if he had been standing a foot to the left, as suggested by the account of Moss (2012). If the foreman can't rule out that salient possibility, Moss would say, it would be irresponsible of him to claim that the worker would have been killed if he had been standing a foot to the left.

If *if*-clauses can 'live on' in discourse beyond their local domain, there should be nothing preventing them from pragmatically restricting non-local nominal domains.

There is a solution for Abbott's Puzzle, then. The solution says that when we interpret 14(a) or (b), the domains of the embedding quantifier phrases are pragmatically restricted by the embedded *if*-clause. The embedded conditional itself might very well be Hook.

4.5 The Family of Hook

Hook is just one among many kinds of conditionals that can be embedded under quantifiers. All of the following pairs of conditionals feel equivalent.

- (29) a. Everyone failed if they goofed off.
b. Nobody passed if they goofed off.
- (30) a. Everyone will fail if they goof off.
b. Nobody will pass if they goof off.
- (31) a. Everyone would fail if they goofed off.
b. Nobody would pass if they goofed off.
- (32) a. Everyone should fail if they goof off.
b. Nobody should pass if they goof off.
- (33) a. Everyone has to fail if they goof off.
b. Nobody can pass if they goof off.

By the end of the day, we would want to account for all the equivalences in (29) to (33). We would also want to explain what is going on with 34(a) and (b).

- (34) a. Everyone is likely to fail if they goof off.
b. Nobody is likely to pass if they goof off.

34(b) seems to have two interpretations. One, but not the other, makes 34(a) and (b) equivalent. Suppose everyone who goofs off has a 50% chance of passing. Then 34(a) is false. On its first interpretation, 34(b) says that everyone is unlikely to pass if they goof off. That's also false if everyone who goofs off still has a 50% chance of passing. On its second interpretation, 34(b) is true on our scenario. Students who have a mere 50% chance of passing if they goof off cannot be said to be likely to pass. None of those who goof off are among those who are likely to pass, then. 34(a) and (b) are equivalent on the first interpretation, but not on the second.

We need to steer a course that allows the whole family of conditionals to stand united. The conditionals embedded in (29) to (34) are all different, but they don't differ in capricious ways. The main difference is the modal in their consequent: *likely*, *can*, *have to*, *should*, *would*, and *will*. If we want a unified analysis of all *if*-clauses in (30) to (34), we need to let the *if*-clauses restrict the domains of their modals. This is the Restrictor View of *if*-clauses. There is also the apparently modal-less (29). For full generality, we should posit a silent modal in (29). I will use the symbol $\Box$ for that particular modal in what follows. An immediate consequence of the Restrictor View is that the negation of a conditional should amount to the negation of its (restricted) modalized consequent. This is a particular interesting prediction to check, since different types of modals are known to interact with negation in different ways:

- (35) a. Nobody will fail.
 b. Nobody would fail.
 c. Nobody should fail.
 d. Nobody has to fail.
 e. Nobody can fail.
 f. Nobody is likely to fail.

The interaction between negation and conditionals has recently been investigated experimentally by Paul Égré and Guy Politzer (reported in Égré and Politzer, 2013). Preliminary findings suggest that conditionals might interact with negation in the way expected on the Restrictor View.

If the *if*-clauses in (29) to (34) restrict their modal, they simultaneously restrict two domains: that of the modal and that of the quantifier. The restriction of the quantifier is a pragmatic effect, as we saw earlier. The restriction of the modal seems to be more tightly engineered by grammar. To see the difference, we need to move to a slightly more technical level of discussion. For illustration, I will adopt von Fintel's (1994) implementation of the Restrictor View.

Suppose every occurrence of conditional *if* carries a domain variable that ranges over accessibility functions (functions from worlds to sets of worlds) and is coindexed with a domain variable on a local modal.

- (36) If_{c₁} she goofed off, she has to_{c₁} fail.

The interpretations of *if*, *has to*, and *can* could be as in (37).

- (37) For any sentences α and β :

- a. $[[[If_{c_1}\beta]\alpha]]^{w,g} = [[\alpha]]^{w,g'}$, where g' is like g , except that $g'(C_I) = \lambda w. g(C_I)(w) \cap [[\beta]]^g$.
- b. $[[has\ to_{c_1}\alpha]]^{w,g} = 1$ iff $g(C_I)(w) \subseteq [[\alpha]]^g$.
- c. $[[can_{c_1}\alpha]]^{w,g} = 1$ iff $g(C_I)(w) \cap [[\alpha]]^g \neq \emptyset$.
- d. $[[\alpha]]^g =_{\text{def}} \{w: [[\alpha]]^{w,g} = 1\}$.

If-clauses do not identify the values of the domain variables they are coindexed with on von Fintel's account. They merely constrain their values. We might assume that initial variable assignments assign the trivial accessibility function (the function that assigns the set of all possible worlds to every world) to each modal domain variable. Context and *if*-clauses can then successively update the values of those variables. The account assumes that *if*-clauses are coindexed with a local modal contained in the sentence they are adjoined to. The coindexation has the effect that the domain of the *if*-clause and the domain of the modal are identified. Relying on related observations in Iatridou (1991), von Fintel points out that the locality requirement for the association of an *if*-clause with its modal is the one familiar from overt movement. The relation between *if*-clause and its modal can be long-distance, but it cannot be across known barriers for movement. This suggests that by the time we see an *if*-clause, it may have moved away from its original position adjacent to its modal. Alternatively, it may not be the *if*-clause itself that enters a relation with a modal, but its 'correlate' pronoun *then*. It would then not be the *if*-clause, but its correlate that moves away from an adjacent modal

(see Bhatt and Pancheva, 2006) for discussion of such a possibility within a different analysis of conditionals).

von Fintel's implementation of the Restrictor View predicts that multiple *if*-clauses should be able to restrict one and the same modal, and a single *if*-clause should be able to restrict multiple modals. Both predictions are borne out, as illustrated in (38) and (39).

- (38) a. If he left at five, he must be home, if he didn't stop for a beer.
 b. (If_{c₁} he left at five (If_{c₁} he didn't stop for a beer (he has to_{c₁} be home)))
 (39) If_{c₁} a wolf entered the house, he must_{c₁} have eaten grandma, since she was bedridden. He might_{c₁} have eaten the girl with the red cap, too.

If *if*-clauses are grammatically required to relate to a modal, but modals are not grammatically required to relate to an *if*-clause, there is a grammatically enforced relation between *if* and *must* in (39), but there is no grammatically enforced relation between *if* and *might*, nor between *might* and *must*. There is nothing in the grammar that requires coindexation of the domain variables of the two modals. The domain variables of modals can be restricted by context alone, as in (40):

- (40) There might_{c₁} be a storm. We might_{c₂} be without electricity.

One way of understanding the second sentence of (40) is as conveying that we might be without electricity if there is a storm. The first sentence in (40) raises the possibility that there might be a storm. As a result, the possible worlds considered for the second sentence can be pragmatically restricted to those where there is a storm. Mauri and van der Auwera (forthcoming) report that there are languages where conditionals are generally expressed via a pragmatic strategy along the lines of (40).

On the Restrictor View, the modals in (29) to (34) are the crucial players. They should be responsible for the distinctive properties of all conditionals. I will not be able to demonstrate this in full detail. I would have to dig deeply into the semantics of each individual modal. To illustrate the agenda, I will derive the interpretations of (41) and (42) by positing particular denotations for *will* and the unpronounced modal $\Box$.

- (41) No_D one λx (If_{c₁} x goofs off (will_{c₁} x pass))
 (42) No_D one λx (If_{c₁} x goofed off ($\Box$ _{c₁} x passed))

I am aiming for interpretations where (41), but not (42), cares about what would happen if a student goofed off. On the intended readings, (41) is a generic conditional, and (42) is a 'one-case' conditional that makes a hypothesis about a particular exam.⁵ I am aware that those interpretations do not only depend on the modals. The contributions of tense-mood-aspect marking in the participating sentences are very

⁵ The term 'one-case-conditional' is due to Kadmon (1992). Not all conditionals with silent modals are 'one-case' conditionals. Not all silent modals are $\Box$. Some seem to be genuine necessity operators. The kind of experiments designed by Égré and Politzer (2013) should help with identifying the nature of silent modals in conditionals.

important, too (see Iatridou's contribution in this volume). I will not be able to separate out those contributions from those of the modals themselves in this chapter. That's a project for another time. Here are the proposed denotations of the two modals:

- (43) a. $[[will_c\alpha]]^{w,g} = 1$ iff $[[\alpha]]^{w',g} = 1$, where w' is the world in $g(C)(w)$ that is closest to w . Pick the absurd world if there is none.
 b. $[[\Box_c\alpha]]^{w,g} = 1$ iff $g(C)(w) \cap \{w\} \subseteq [[[\alpha]]^g$.

In interaction with a coindexed *if*-clause, 43(a) produces the Stalnaker conditional, 43(b) delivers the material conditional. Under the assumption that the antecedent is possible, the Stalnaker conditional accounts for the intended interpretation and the equivalence of 30(a) and (b), the pair we started out with.

- (30) a. Everyone will fail if they goof off.
 b. No one will pass if they goof off.

To account for the equivalence and the intended interpretation of 29(a) and (b), the embedded conditional has to be material implication, and the *if*-clause has to pragmatically restrict the domain of the embedding quantifier.

- (29) a. Everyone failed if they goofed off.
 b. No one passed if they goofed off.

If the restriction of nominal domain variables by *if*-clauses is pragmatic, it should be optional in the absence of non-grammatical pressures. Why is it nevertheless obligatory, at least in 29(b)? I already suggested a possible answer for deviant examples like (44):

- (44) #Exactly five queries were answered if they came from a respectable address.

If the presence of the modal $\Box$ makes the embedded conditional a material conditional in (44), we are forced to count queries that came from dubious addresses as confirming instances. I suggested that Nicod's Criterion militates against this. If numeral quantifiers like *exactly five* have no domain variables, the sentence should feel odd. In (29), the violation of Nicod's Criterion can be avoided by letting the *if*-clause restrict the domain variables introduced by the quantifiers. This has an effect on the interpretation of 29(b), but results in an equivalent interpretation for 29(a).

In 30(b), the modal *will* tells us that the embedded conditional is a Stalnaker conditional. It forces us to move on to merely possible worlds and check whether students who don't actually goof off would fail if they did. There is no violation of Nicod's Criterion, then, and no need for the *if*-clause to restrict the nominal domain variable introduced by the quantifier. It should nevertheless be able to. It seems it is: there seems to be a reading of 29(b) that doesn't make Leslie's Meadow (who doesn't goof off, but would pass no matter what) a counterexample.⁶ Here is a slightly

⁶ Thanks to Daniel Rothschild for pointing this out.

different example that brings out that kind of reading more clearly. Suppose that, as a matter of policy, students don't pass a course in Meadow's school if they have skipped more than three classes. Meadow doesn't ever skip classes, but, being the teacher's favorite, she would pass even if she did. There is an interpretation of (45) that feels true on this scenario.

(45) No one will pass if they skipped more than three classes.

If *will* tells us that the embedded conditional in (45) is a Stalnaker conditional, then Meadow would be a counterexample for (45) if the *if*-clause couldn't restrict the domain variable of *no*. We would have to consider the closest world where she did skip classes. She would still pass, and as a consequence, (45) would wind up false. Since (45) has an interpretation where it is true on our scenario, I conclude that the *if*-clause in (45) can optionally restrict the nominal quantifier.

4.6 Outlook

I have scrutinized one particular species of embedded conditionals: those embedded under nominal quantifiers. The display in (29) to (34) gave a snapshot of representative specimens. If we care about a unified treatment of the species, we can't semantically compose the meanings of the sentences in (29) to (34) from just two pieces: the quantifier phrase and, say, a Stalnaker conditional. But we also can't seem to semantically compose their meanings from just three pieces: the quantifier phrase, the *if*-clause, and a modal. There is another force to reckon with. The *if*-clause can pragmatically restrict the domain of the nominal quantifier. Do we have an account of those constructions, then? If so, is it a 'general algorithmic' account? What if not? Is that a reason to just give up?

References

- Abbott, B. (2004) Some remarks on indicative conditionals. Paper presented at Semantics and Linguistic Theory XIV, 2004.
- Bhatt, R. and Pancheva, R. (2006) Conditionals. In M. Everaert and H. v. Riemsdijk (eds), *The Blackwell Companion to Syntax*, Malden, MA: Blackwell, pp. 638–87.
- Dekker, P. (2001) On 'if' and 'only'. Paper presented at Semantics and Linguistic Theory XI, 2001.
- Égré, P. and Cozic, M. (2008) Introduction to the logic of conditionals. ESSLI, University of Hamburg. http://paulegre.free.fr/Teaching/ESSLI_2008/cond_essli08.pdf
- Égré, P. and Politzer, G. (2013) On the negation of indicative conditionals. In M. Aloni, M. Franke, and F. Roelofsen (eds), *Proceedings of the 19th Amsterdam Colloquium*, 10–18. http://www.ilc.uva.nl/AC/AC2013/uploaded_files/inlineitem/02_Egre_Politzer.pdf
- Edgington, D. (1995) On conditionals. *Mind*, New Series, 104 (414): 235–329.
- Edgington, D. (2000) General conditional statements: A response to Kölbel. *Mind*, New Series, 109 (433): 109–16.
- Edgington, D. (2001) Conditionals. In E. N. Zalta (ed.), *Stanford Encyclopedia*. Stanford: CSLI.
- Elbourne, P. (2013) *Definite Descriptions*. Oxford: Oxford University Press.
- Fintel, K. von (1994) *Restrictions on Quantifier Domains*. University of Massachusetts Ph.D.
- Fintel, K. von (1998) Quantifiers and 'if'-clauses. *Philosophical Quarterly* 48: 209–14.

- Fintel, K. von (1999) NPI licensing, Strawson entailment, and context dependency. *Journal of Semantics* 16: 97–148.
- Fintel, K. von (2001) Counterfactuals in a dynamic context. In M. Kenstowicz (ed.), *Ken Hale: A Life in Language*, Cambridge, MA: MIT Press, pp. 123–52.
- Fintel, K. von and Iatridou, S. (2002) If and when ‘if’-clauses can restrict quantifiers. Manuscript, MIT.
- Frank, A. (1996) *Context Dependence in Modal Constructions*. University of Stuttgart PhD dissertation.
- Geis, M. and Zwicky, A. (1971) On invited inferences. *Linguistic Inquiry* 2: 561–6.
- Gillies, T. (2007) Counterfactual scorekeeping. *Linguistics and Philosophy* 30: 329–60.
- Goodman, N. 1955. *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.
- Higginbotham, J. (1986) Linguistic theory and Davidson’s program in semantics. In E. Lepore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Oxford: Basil Blackwell, pp. 29–48.
- Higginbotham, J. (2003) Conditionals and compositionality. *Philosophical Perspectives* 17: 181–94.
- Huitink, J. (2009) Quantified conditionals and compositionality. *Language and Linguistics Compass* 4: 42–53.
- Iatridou, S. (1991) Topics in Conditionals. Cambridge, MA: MIT Ph.D. dissertation.
- Iatridou, S. (2019) Grammar matters. Chapter 8, this volume.
- Kadmon, N. (1992) Unique and non-unique reference. *Outstanding Dissertations in Linguistics*. New York: Garland.
- Keshet, E. (2008) *Good Intensions: Paving Two Roads to a Theory of the de re/de Dicto Distinction*. MIT PhD dissertation.
- Klinedinst, N. (2011) Quantified conditionals and conditional excluded middle. *Journal of Semantics* 28: 149–70.
- Kölbel, M. (2000) Edgington on compounds of conditionals. *Mind*, New Series, 109 (433), 97–108.
- Landman, F. (2004) *Indefinites and the Type of Sets*. New York: Wiley-Blackwell.
- Leslie, S.-J. (2009) ‘If’, ‘unless’, and quantification. In R. J. Stainton and C. Viger (eds), *Compositionality, Context and Semantic Values: Essays in Honor of Ernie Lepore*, Dordrecht: Springer, pp. 3–30.
- Lewis, D. K. (1976) Probabilities of conditionals and conditional probabilities. *Philosophical Review* 85: 297–315.
- Martí, L. (2003) *Contextual Variables*. University of Connecticut PhD dissertation.
- Mauri, C. and van der Auwera, J. (forthcoming). Chapter 19: Connectives. In K. M. Jaszczolt and K. Allan (eds), *Cambridge Handbook of Pragmatics*, Cambridge: Cambridge University Press.
- McGee, V. (1985) A counterexample to modus ponens. *Journal of Philosophy* 9: 462–71.
- Moss, S. (2012) On the pragmatics of counterfactuals. *Nous* 46: 561–86.
- Nicod, J. (1924) *Le Problème Logique de l’Induction*. Paris: Félix Alcan.
- Pizzi, C. and Williamson, T. (2005) Conditional excluded middle in systems of consequential implication. *Journal of Philosophical Logic* 34: 333–62.
- Schwarz, F. (2009) *Two Types of Definites in Natural Languages*. University of Massachusetts at Amherst PhD dissertation.
- Schwarz, F. (2012) Situation pronouns in determiner phrases. *Natural Language Semantics* 20: 431–75.

- Stalnaker, R. (1968) A theory of conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, Vol. 2: American Philosophical Quarterly Monograph Series, Oxford: Basil Blackwell, pp. 98–112.
- Stalnaker, R. (1975) Indicative conditionals. *Philosophia* 5: 269–86.
- Stalnaker, R. (2014) *Context*. Oxford: Oxford University Press.
- Stanley, J. (2000) Context and logical form. *Linguistics and Philosophy* 23: 391–434.
- Stanley, J. and Szabó, Z. G. (2000) On quantifier domain restriction. *Mind & Language* 15: 219–61.

New Paradigm Psychology of Conditional Reasoning and Its Philosophical Sources

David Over

The psychology of reasoning is going through a paradigm shift in the way it studies human reasoning (Elqayam and Over, 2013; Manktelow, Over, and Elqayam, 2011; Over, 2009). The older, *binary and extensional paradigm* in the psychology of reasoning was mainly focused on deductive inference from assumptions. Participants in experiments were asked to assume premises given to them, excluding or setting aside any relevant beliefs they might have, and to draw only conclusions that necessarily followed. The premises and conclusion were only classified as true or false, and the inference was only classified as logically valid or logically invalid. In the most influential psychological theory in the old paradigm, the normative standard for ‘correct’ logical inference was extensional logic, and the natural language indicative conditional was equivalent to the truth functional conditional (Johnson-Laird and Byrne, 1991). If the participants endorsed ‘valid’ inferences according to this standard, they were deemed to give ‘correct’ answers. If they did not do this, but introduced relevant beliefs into their inferences and reasoned probabilistically, they were ‘biased’ and committed ‘fallacies’. In the *new Bayesian/probabilistic paradigm*, theorists recognize the uncertainty that is usually present in the premises of even deductive inferences, and they reject the truth functional interpretation of the indicative conditional (Evans and Over, 2004; Oaksford and Chater, 2007, 2009; Pfeifer and Kleiter, 2010; Politzer, Over, and Baratgin, 2010). The philosophical sources of this new approach are clear in its citations: from de Finetti (1936/1995, 1937/1964) and Ramsey (1926/1990) to Adams (1975, 1998) and Edgington (1995, 2003).

Most useful inferences, in both science and everyday affairs, are from uncertain beliefs or plausible hypotheses that are taken to be relevant to a matter at hand.

* I would like to thank Dorothy Edgington for many years of illuminating discussions about conditionals, since I was her first PhD student in the early 1970s, and continuing long after I took up the psychology of reasoning. I would also like to thank Lee Walters for helpful comments and corrections on my first draft, and the participants at the 2011 conference in Dorothy’s honour at the Institute of Philosophy, London, for their comments on my talk, which formed the basis of this chapter.

Consider an example of a major premise for the valid inference form of *Modus Ponens* (MP), inferring q from p and *if p then q* :

- (1) If global warming continues (g), then London will be flooded (l).

We cannot simply assume *if g then l* as the major premise, and g as the minor premise, of MP when we are trying to predict the future or make a rational decision about reinforcing the Thames Barrier, which protects London from flooding. Uncertainty in the premises of even a deductive inference should affect confidence in the conclusion, and of course this point has long been recognized in logic and philosophy. People are usually uncertain in their beliefs to some degree, even if they are overconfident, and this uncertainty should affect confidence in the conclusions that they infer from their beliefs. In psychological experiments on reasoning, uncertainty in the premises of deductive inferences has been found to affect confidence in the conclusions (Politzer, 2005; Stevenson and Over, 1995, 2001), and theories of reasoning in the new paradigm aim to explain this fact (Evans and Over, 2004; Oaksford and Chater, 2007; Pfeifer and Kleiter, 2010).

The most prominent account of reasoning in the old extensional and binary paradigm was the *mental model theory* of Johnson-Laird and Byrne (1991). Their commitment to a binary and extensional theory is seen most clearly in their account of natural language indicative conditionals like (1). They argued (pp. 7 and 74) for the truth functional analysis of this kind of conditional, which makes *if p then q* logically equivalent to the truth functional, *material conditional*, logically equivalent to *not- p or q* . There could not be a better illustration of commitment to the binary paradigm than their argument at this point. After noting that *if p then q* is true when p is true and q is true, and false when p is true and q is false, they asked (about their example of an indicative conditional) whether *if p then q* is true or false when p is false. They replied, 'It can hardly be false, and so, since the propositional calculus allows only truth or falsity, it must be true' (p. 7). They did not ask whether the conditional might be non-truth functional, or whether it might be neither true nor false when its antecedent is false. Yet it was known back then that participants in truth table experiments do not classify *if p then q* as true or false when p is false, but rather say, in a so-called *defective truth table*, that the falsity of p is 'irrelevant' to the truth or falsity of *if p then q* (see Evans and Over, 2004, for early research on truth table responses in these experiments).

Johnson-Laird and Byrne (1991) argued explicitly for the logical validity of the paradoxes of claiming that the indicative conditional of natural language is the truth functional material conditional. The two basic paradoxical inferences are inferring *if p then q* from *not- p* alone and from q alone. They claimed that people only fail to endorse these inferences for pragmatic reasons (pp. 73–4). Byrne and Johnson-Laird (2009) argued that it is 'correct' to say that the probability of the natural language indicative conditional $P(\text{if not-}p \text{ then } q)$ is the probability of the disjunction, $P(p \text{ or } q)$. They also argued that it is logically valid to infer *if not- p then q* from p or q , but at the same time, they inconsistently denied that their account of the conditional makes it truth functional (see also Johnson-Laird and Byrne, 2002; for critical replies see Evans and Over, 2009; Over, Evans, and Elqayam, 2010; and Politzer et al., 2010). A reading of Edgington (1995) would have shown what is wrong with claiming that the paradoxes are valid, that $P(\text{if not-}p \text{ then } q) = P(p \text{ or } q)$, and that the inference from p or q to *if not- p then q* is valid. It would also have prevented the inconsistency

of taking these positions while claiming that the account given of the conditional is not truth functional.

5.1 The Equation and the de Finetti Table

Fortunately, there were other psychologists who were influenced by Edgington's arguments that the paradoxes are invalid, and that the subjective probability of a natural language conditional *if p then q* is the subjective conditional probability, $P(q|p)$. She expressed the latter view in what she called *The Equation* (see Douven, 2008, 2016, and Kaufmann, 2004, for the normative debate). Psychologists have turned this into a descriptive claim and tested it as *the conditional probability hypothesis*:

$$P(\text{if } p \text{ then } q) = P(q|p)$$

There is very strong support in experiments for the conditional probability hypothesis across a wide range of indicative conditionals (see Skovgaard-Olsen, Singmann, and Klauer, 2016; and Over and Cruz, 2018, for possible qualifications), but none for the truth functional representation of the indicative conditional, as found in mental model theory (Byrne and Johnson-Laird, 2009; Johnson-Laird and Byrne, 1991, 2002). Actually, this theory only presents a precise mental model for the so-called *basic conditional*, the content of which is unaffected by context and background knowledge, with an antecedent and consequent that are semantically independent of each other (Johnson-Laird and Byrne, 2002; Evans and Over, 2004). An example of a 'typical' basic conditional was said to be, *if the weather is fine then the sun shines*. A basic conditional *if p then q* is supposed to have an initial mental model that is equivalent to the conjunction, $p \ \& \ q$, plus a 'mental footnote', indicated by three dots, that there are further implicit models:

p	q
...	

People are also supposed to be able to 'flesh out', or fully represent, all the mental models for a conditional:

p	q
not-p	q
not-p	not-q

The above models are equivalent to the representation of the material conditional, *not-p or q*, and the three lines of the binary truth table that make *not-p or q* true. Two predictions follow from mental model theory about probability judgments. People who have only the initial model of the conditional will judge $P(\text{if } p \text{ then } q) = P(p \ \& \ q)$, and people with the full models of the conditional will judge $P(\text{if } p \text{ then } q) = P(\text{not-}p \text{ or } q)$.

The first experiments comparing the conditional probability hypothesis with the predictions of mental model theory asked participants to make a probability judgment about a 'basic' *singular* conditional (Evans et al., 2003: 333), referring to a single object randomly selected from a given frequency distribution (see also Oberauer and Wilhelm, 2003, for somewhat different experiments, and Cruz and Oberauer, 2014,

on general conditionals). In experiments of this type, most participants judge that $P(\text{if } p \text{ then } q) = P(q|p)$, confirming the conditional probability hypothesis. We can illustrate this finding with the experiment in Politzer et al. (2010), which also extended the results to judgments about conditional bets. They gave participants a drawing of seven chips. There were two black circular chips, one white circular chip, three black square chips, and one white square chip. The participants were told that one chip was going to be drawn from this distribution of chips. One group was then given a singular indicative conditional about the random chip:

- (2) If the chip is square (s), then it will be black (b).

Another group was given a conditional bet between two players, Mary and Peter, about the random chip:

- (3) I bet you 1 Euro that if the chip is square (s), then it will be black (b).

The first group was asked for the probability that (2) was true, and the second group for the probability that Mary would win her bet (3). The point of the experiment was to test whether there is a parallel relationship between the indicative conditional (2), the conditional bet (3), and the conditional probability, as implied by both de Finetti (1936/1995, 1937/1964) and Ramsey (1926/1990, 1929/1990). By this hypothesis, we should find the following. The indicative conditional (2) is true when the random chip is square and black, is false when it is square and white, and is 'void', neither true nor false, when the chip is circular. Mary's conditional bet (3) is won when the random chip is square and black, is lost when the chip is square and white, and is 'void', called off and neither won nor lost, when the chip is circular. The probability that the assertion (2) is true is the probability of the true case, s & b , given that an indicative assertion (the void case does not occur) is made: $P(s \& b|s) = P(b|s)$. The probability that the bet (3) is won is the probability of the winning case, s & b , given that a bet (the void case does not occur) is made: $P(s \& b|s) = P(b|s)$. In other words, the void cases are ignored as irrelevant to a question about the probability of the truth of a conditional assertion or the probability of winning a conditional bet, for the reason that there is no indicative assertion or bet in the void cases. Another example is a conditional promise, such as when a father promises his daughter, 'If you wash the car (w) I will give you 10 Euros (t).' The probability that the father will keep his promise is $P(t|w)$, and not $P(w \& t)$, because the question presupposes that there is a promise, making the void case, of the daughter not washing the car, irrelevant.

The majority of participants, in the experiment of Politzer et al. (2010), responded that the probability of the truth of (2) is $P(b|s)$, and that the probability the bet (3) is won is $P(b|s)$. The participants had a drawing of the seven chips right in front of them, and it would have been easy for them to 'flesh out' all the mental models for (2) and (3), if they used mental models at all. And yet going against mental model theory, very few participants gave $P(\text{not-}s \text{ or } b)$ as these probabilities. Also confirming the hypothesis of a parallel relationship, participants said that a square and black chip, s & b , made (2) true and won the bet (3), that a square and white chip, s & $\text{not-}b$, made (2) false and lost the bet (3), and that (2) was neither true or false, and the bet (3) neither won nor lost, when the random chip was circular, $\text{not-}s$, whether it was white or black.

Table 5.1. The ‘defective’, 2×2
de Finetti table for *if p then q*

<i>q</i>	1	0
<i>p</i>		
1	1	0
0	V	V

1 = true, 0 = false, and V = void.

The results on conditional bets were new, but as noted above, it was already known that participants respond with a three-valued, ‘defective’ truth table in truth table tasks (Evans and Over, 2004). They will say the *s* & *b* case makes *if s then b* true, the *s* & *not-b* case makes *if s then b* false, and the *not-s* cases are ‘irrelevant’ to the truth or falsity *if s then b*. According to mental model theory, the ‘correct’ response is supposedly the full binary truth table for the material conditional, *not-s or b*, and that implies of course that the falsity of *s* makes (2) true and wins the conditional bet (3) for Mary. However, there is nothing ‘defective’ in the negative sense about the three-valued ‘defective’ table from the point of view of the new paradigm, and it should be called the *2x2 de Finetti table* after de Finetti, who first proposed it on normative grounds. In this table, the third value is the ‘void’ case, when the indicative conditional is neither true nor false, and the conditional bet is neither won nor lost (Baratgin, Over, and Politzer, 2013; de Finetti, 1936/1995, 1937/1964; Politzer et al., 2010). See Table 5.1 for the ‘defective’, 2x2 de Finetti table.

A further point about this table that Edgington (2003) makes is that it does not represent ‘pleonastic’ uses of ‘true’ (see also Adams, 1998, on pragmatic uses of ‘true’). People sometimes use ‘true’ simply to endorse an assertion, and they can use it in this way for a conditional *if p then q* without implying that the indicative, or factual, state of affairs holds in which both *p* and *q* are true. For example, people might say that *if s then s* is ‘certainly true’ without implying that *s* is true. The use of ‘true’ in ‘certainly true’ is pleonastic, and people are only saying that *if s then s* is ‘certain’, $P(s|s) = 1$.

In experiments on frequency distributions like the one we have just described, there tends to be a minority who judge the probability of the truth of the conditional *if p then q* to be probability of the conjunction *p* & *q*, $P(\text{if } p \text{ then } q) = P(p \& q)$. Politzer et al. (2010) found a similar minority conjunctive response to the question about the probability of winning the conditional bet. It could be claimed that these results confirm the prediction of mental model theory that some participants have an initial model for the conditional that is equivalent to *p* & *q*. However, Fugard, Pfeifer, Mayerhofer, and Kleiter (2011) showed that people tend to switch from the conjunctive to the conditional probability response as they do more and more tasks in which they are asked for probability judgments about conditionals referring to frequency distributions. In experiments on ‘causal’ conditionals like (1), which are not about an object selected from an artificially or abstractly given frequency distribution, the conditional probability response is dominant, and there is little evidence for the conjunctive probability response (Over, Hadjichristidis, Evans, Handley, and Sloman, 2007).

Evans et al. (2003) and Evans and Over (2004) hypothesized that the conjunctive response, in experiments where abstract frequency distributions are used, is due to a processing problem. The conditional probability response comes from full processing of the conditional. This hypothesis has been confirmed in more recent experiments. Participants who give the conjunctive response had relatively low cognitive ability, and those who give the conditional probability response had relatively high cognitive ability. These high ability, conditional probability responders also tended to be those who responded with the 'defective', de Finetti table (Evans, Handley, Neilens, and Over, 2007; Politzer et al., 2010). If mental model theory were correct, high cognitive ability participants would be able to make all the mental models explicit, and so judge $P(\text{if } p \text{ then } q) = P(\text{not-}p \text{ or } q)$ and respond with the full binary truth table for *not- p or q* . There is also a developmental trend, in which young children respond with the conjunctive probability, but this response declines with age, and the conditional probability judgment increases until it is the majority response in undergraduate students (Gauffroy and Barrouillet, 2009).

Giroto and Johnson-Laird (2004) claimed that the conditional probability response is the result of only one 'strategy' that people will use to judge the probability of a conditional, with another significant 'strategy' resulting in the conjunctive response. But they cannot explain why people tend to switch from the conjunctive to the conditional probability response as they get more and more practice in experiments about abstract frequency distributions (Fugard et al., 2011). Their claim was that people judge $P(\text{if } p \text{ then } q) = P(q|p)$ because they wrongly interpret a question about the probability of the whole conditional—the *wide scope* use of the probability operator—as if it were a question about the probability of the consequent if the antecedent holds—the *narrow scope* use of the probability operator. They held that, when people respond $P(\text{if } p \text{ then } q) = x$, what they mean is that *if p then $(P(q) = x)$* . But in fact, every effort has been made in the written materials of almost all experiments on the conditional probability hypothesis, from Evans et al. (2003) to Politzer et al. (2010) and Fugard et al (2011), to ensure that the participants interpret the scope of the probability operator in the conditionals as wide, as applying to the whole conditional.

Giroto and Johnson-Laird presupposed that the probability operator is given narrow scope by people when they use conditionals like, *if the chip is square then it is probably black*. Yet linguists and logicians would point out that the probability modal, like other modals, can cause scope ambiguities, with it being unclear whether the scope of the modal is wide or narrow. The probability modal in the conditional, *if the chip is square then it is probably black*, can be given narrow scope, *if the chip is square then (it is probably black)*, but also wide scope, *it is probable that (if the chip is square then it is black)*. Over, Douven, and Verbrugge (2013) ran experiments directly on such scope ambiguities and confirmed that people interpret the probability operator as having wide, and not narrow, scope.

Giroto and Johnson-Laird themselves committed a modal fallacy in interpreting their results (Over et al., 2013; Over et al., 2007; Milne, 2012; Politzer et al., 2010), and Byrne and Johnson-Laird (2010) went so far as to imply that wide and narrow scope interpretations of modals are synonymous. They specifically claimed that the necessity modal applied to the whole conditional, wide scope, 'is synonymous with' this

modal applied only to the consequent, narrow scope (p. 56). They stated that *It is necessary that (if A then B)*, wide scope, is synonymous with *if A then (it is necessary that B)*, narrow scope. This claim implies that all truth is necessary truth: that a statement is true if and only if it is necessarily true (Politzer et al., 2010). It is a notorious fallacy in logic and philosophy, used in sophistical ‘proofs’ of logical determinism (Swartz and Bradley, 1979). The equivalent fallacy for the probability modal implies that all true statements have a probability of 1. Yet again a reading of Edgington (1995) would have prevented these fallacies.

There is a clear difference between, for instance, *having absolute confidence in (if p then p)* and *if p then (having absolute confidence in p)*, the latter expressing omniscience, just as *if A then (we know A)* does. Theorists can get into a muddle and commit fallacies when they first start to think about modality and scope. But how do we test ordinary people’s interpretation of the scope of a modal in a psychological experiment? How do we know whether they are using *if p then probably q* to mean the wide scope *probably (if p then q)* or the narrow scope *if p then (probably q)*? Over et al. (2013) use conditional inferences for this purpose. Consider *Modus Tollens* (MT): inferring *not-p* from the major premise *if p then q* and the minor premise *not-q*. With *not-(probably q)* as the minor premise, MT can be applied assuming that the modal has narrow scope, *if p then (probably q)*, but not assuming that it has wide scope, *probably (if p then q)*. Assuming that the scope is narrow, *if p then (probably q)*, and *not-(probably q)* is the minor premise, MT allows us to infer validly that *not-p* definitely holds. But this inference is invalid for the wide scope interpretation, *probably (if p then q)*. Over et al. (2013) find that the endorsement rate for MT falls when participants are given *if p then probably q* and *not-(probably q)* as premises, showing that participants take the scope as wide and not narrow. This result further supports the findings that people are making a probability judgment about the entire conditional when they respond that $P(\text{if } p \text{ then } q) = P(q|p)$ in experiments on the conditional probability hypothesis.

There is thus substantial evidence that people understand the natural language indicative conditional to be such that The Equation holds for it. A conditional *if p then q* with the property that $P(\text{if } p \text{ then } q) = P(q|p)$ has been called the *probability conditional* (Adams, 1998) and also the *conditional event* (de Finetti, 1937/1964; Gilio, 2002). By the de Finetti table, such a conditional is ‘void’ when its antecedent is false and so *not-p* holds (Politzer et al., 2010). Suppose that (1) was understood to be a prediction about the next ten years, but that over this time environmental measures have stopped global warming, and London is not flooded. The de Finetti table, as a psychological hypothesis, implies that we would tend to lose interest in (1) as we became more and more confident that the antecedent of this indicative conditional was false. We might then become more inclined to assert a counterfactual conditional, which differs from the indicative in the presupposition that its antecedent is false. Suppose we finally become sure that the antecedent of (1) is false about the last ten years. We will now consider (1) ‘void’. We will not even use the present or past tense form of (1), but will assert outright the counterfactual, *if global warming had continued then London would have been flooded*. There is little work as yet on the probability of counterfactuals in the new paradigm, but what there is does support a relation between counterfactuals and The Equation (Over et al., 2007). We will,

however, restrict our attention here to indicative conditionals (see Edgington, 2008, 2014; Over, 2017; and Over and Hadjichristidis, 2009, on counterfactuals).

5.2 Extending the de Finetti Table and the Ramsey Test

It is important to be clear about what ‘void’ means for indicative conditionals with false antecedents. Edgington (1995) criticizes Quine (1952) for claiming that it is ‘as if nothing has been said’ by *if p then q* when *p* turns out to be false. Her example is that uttering, *if you press that switch then there will be an explosion*, might bring it about that the switch was not pressed, and that could hardly mean in turn that there had been nothing in the utterance. Similarly, reflecting on conditionals like (1) might motivate us to take steps to stop global warming, but for that very reason, something significant was still expressed, which could then be conveyed, as we have just noted, by the counterfactual, *if global warming had continued then London would have been flooded*. To hold that the indicative is ‘void’ because *p* is false simply means that no indicative and categorical fact holds: neither *p* & *q* holds nor *p* & *not-q* holds. But there is still the conditional probability, $P(q|p)$, and we can follow Jeffrey (1991a) in thinking of $P(q|p)$ itself as a kind of ‘truth value’ for *if p then q* when *p* is false. It is a ‘value’ unlike the other two, truth and falsity, in being epistemic and subjective (Edgington, 1995; Over and Cruz, 2018; and see also Edgington, 2004, on objective probabilities and conditionals).

We have seen above that the ‘defective’ truth table is better termed the ‘de Finetti table’ (see again Table 5.1). In traditional psychological experiments on this table, the *p* and *q* components of *if p then q* were always seen by the participants to be true, or seen to be false, and were never themselves uncertain. But from the point of view of the new paradigm, this was a limitation on the classifications that we can make. For example in (1), the antecedent, *g*, and consequent, *l*, are obviously uncertain, to some extent for most of us. To begin to overcome this restriction, Baratgin et al. (2013) extended the experiment of Politzer et al. (2010) on (2), *if s then b*, to make the shape, *s*, and the colour, *b*, of the chip visually uncertain. More technically, the 2x2 truth table was expanded to a 3x3 table. In this 3x3 table, *s* and *b* could be seen to be true, seen to be false, or be visually uncertain with a ‘filter’ used to obscure the chip’s shape or colour. The participants filled in the nine cells for *if s then b* as either *true*, *false*, or *uncertain*.

Logicians and philosophers have proposed a number of 3x3 truth tables for the conditional, and de Finetti himself extended his 2x2 table, Table 5.1, to such a 3x3 table (Baratgin et al., 2013; de Finetti, 1936/1995). See Table 5.2. In a *de Finetti* 3x3

Table 5.2. The 3 × 3 general de Finetti table for *if p then q*

<i>q</i>	1	U	o
<i>p</i>			
1	1	U	o
U	U	U	U
o	U	U	U

1 = true, o = false, and U = uncertain.

Table 5.3. The Jeffrey table for *if p then q*

<i>q</i>	1	0
<i>p</i>		
1	1	0
0	$P(q p)$	$P(q p)$

1 = true, 0 = false, and $P(q|p)$ = the subjective probability of q given p .

table, *if s then b* takes the value of uncertainty whenever s is uncertain or b is uncertain. Another possible 3x3 table is the same as de Finetti's except that *if s then b* is false, rather than uncertain, when s is uncertain and b is false (Farrell, 1979), and there are other distinct tables that have been proposed in the literature on three-valued systems. Nevertheless, Baratgin et al. (2013) find that most participants respond with de Finetti's 3x3 table in this experiment, confirming the hypothesis that his 3x3 table best represents people's responses to uncertainty in the components of a conditional. As they explain, their experimental technique can also be developed to refine 'uncertainty' into different degrees of subjective probability, with the aim of testing whether the 'third' value can ultimately be identified with the conditional probability of b given s , $P(b|s)$, as given in what has been called the *Jeffrey table* (Cruz and Oberauer, 2014; Over and Cruz, 2018). See Table 5.3.

There is then some experimental support for the 3x3 de Finetti table, as well as considerable support for The Equation and the 'defective', 2x2 de Finetti table. All these results are predicted from a new paradigm point of view. Moreover, psychologists in the new paradigm generally agree that people use the procedure that Ramsey (1929/1990) described in a footnote for making probability judgments about conditionals, which has come to be known as the *Ramsey test* (Edgington, 1995; Evans et al., 2003; Evans and Over, 2004; Oaksford and Chater, 2007). Using the Ramsey test (as modified by Stalnaker, 1968), we would fix our degree of belief in (1) by hypothetically supposing that global warming continues, making whatever adjustments might be necessary for consistency, and then judging how likely it is that London will flood under that supposition. To make this judgment, we might use our beliefs about the causal factors that affect sea levels, or we might reply on what we have heard from what we take to be authoritative sources. To make a similar judgment about (2) or (3), we would simply reply on the frequency distribution we are told about. The test would be of an indicative conditional when the antecedent was uncertain. And it would be of a counterfactual when the antecedent was known to be false, making the indicative void.

Some proposals in the logical and philosophical literature, for avoiding the absurdities of the truth functional analysis of ordinary conditionals, make indicative conditionals true or false, and not void, when their antecedents are false. Some of these also try to appeal to the Ramsey test (Douven, 2011; Edgington, 1995; Stalnaker, 1968). However, Lewis (1976) proved the first of a series of *triviality results* for these systems when $P(\text{if } p \text{ then } q) = P(q|p)$, which is now so strongly supported by experiments. One of his ancillary assumptions was indeed that there are no third-value cases, or truth value gaps, for *if p then q*, and triviality results do not extend to the probability conditional. Lewis himself was well aware that there are not triviality results for this

conditional, just as there could not be for conditional bets or conditional promises. However, more research is also needed in the new paradigm on the relation between the growing support for the psychological conditional probability hypothesis and the implications of Lewis's logical proof (Douven and Dietz, 2011).

5.3 Valid Inferences in the New Paradigm

There are many results in psychology showing that people are not always coherent in their probability judgments, including of conditional probability (Barbey and Sloman, 2007). The results of this research do not imply a problem with The Equation or the conditional probability hypothesis, as long as people's judgments about conditionals are as coherent, or incoherent, as their conditional probability judgments. There are also many psychological results implying that people do not always generate or endorse logically valid inferences (Evans and Over, 2004). The new paradigm does not imply that people are never fallacious in their reasoning (but see also Hahn and Oaksford, 2007), and psychology should of course study the fallacies that remain after the right normative standard is applied (see Elqayam and Evans, 2011, for a debate about the place of normative evaluations in psychology). However, the normative relation between coherent probability judgments and logical validity is so close and deep in the new paradigm that the latter can be defined in terms of the former. A single premise inference can be defined as valid if and only if the probability of its premise cannot be coherently greater than the probability of its conclusion (compare Tversky and Kaheman, 1983, on the special case of inferring p from $p \& q$). For inferences with more than one premise, it is helpful to define a special technical sense of 'uncertainty'. Let the *uncertainty*, in this sense, of a statement p be one minus its probability, $1-P(p)$, and of a conditional *if p then q* be $1-P(q|p)$. Now an inference is valid if and only if it cannot increase uncertainty, that is, if and only if the uncertainty of its conclusion cannot be coherently greater than the sum of the uncertainties of its premises (Adams, 1998).

This definition of validity, *probabilistic validity* or *p-validity*, can be used to investigate whether people conform to the validity, or invalidity, of inference forms when the premises are uncertain (Cruz, Baratgin, Oaksford, and Over, 2015; Evans, Thompson, and Over, 2015; Pfeifer and Kleiter, 2010; Singmann, Klauer, and Over, 2014). Such studies give us much more information than binary results, which merely tell us whether or not participants endorse an inference from assumptions (as reported in Johnson-Laird and Byrne, 1991, and other binary paradigm works). For example, consider an instance of MP when the premises are uncertain. Suppose we judge that $P(\text{if } p \text{ then } q) = P(q|p) = 0.6$, and $P(p) = 0.6$. What degree of belief should we have in the conclusion? By elementary probability theory, we know that:

$$P(q) = P(p)P(q|p) + P(\text{not-}p)P(q|\text{not-}p)$$

$$P(q) = 0.36 + 0.4P(q|\text{not-}p)$$

The premises of MP do not supply information about $P(q|\text{not-}p)$, but $P(q|\text{not-}p)$ has a minimum value of 0 and a maximum value of 1. Therefore, the probability of the

conclusion of this MP inference, $P(q)$, should have a minimum value of 0.36 and a maximum value of 0.76. If we are coherent, we should judge that $P(q)$ is somewhere in this interval (Pfeifer and Kleiter, 2010). To take just one example of the limitations of the old binary and extensional paradigm, psychologists working in this paradigm could not even formulate the notion that participants in an experiment could have too much confidence in the conclusion of a valid inference like MP. Participants would have this kind of incoherence, which would be totally missed in the old paradigm, if they judged $P(q) > 0.76$ in an experiment in which $P(\text{if } p \text{ then } q) = P(q|p) = 0.6$, and $P(p) = 0.6$. The new paradigm covers such cases and could potentially connect them with the study of overconfidence in other parts of psychology.

Most conditional inferences studied in the old paradigm were *elimination* inferences, in which the conditional is eliminated and a categorical conclusion inferred. For example, both MP and Affirmation of the Consequent (AC), which is inferring p from *if p then q* and q , are elimination inferences. One problem with old paradigm studies of these inferences is that these studies cannot distinguish between the material conditional and the probability conditional (Over et al., 2010). For both conditionals, MP is valid, and AC is invalid. The new paradigm should turn more of its attention to *introduction* inferences, especially for conditionals (see also Milne, 2012, on the need for a wider range of experiments). One of the simplest introduction inferences for the conditional is inferring *if not- p then q* from p or q , which is valid for the material conditional. It is invalid for the probability conditional, but can often be a probabilistically strong inference for this conditional (Gilio and Over, 2012). It can enable us to infer confidence in a probability conditional, and we can then use MP to infer confidence in the consequent when we become more confident of the antecedent. Studying this inference form could also help us to describe the Ramsey test in detail (Over et al., 2010).

Stalnaker (1975) and Edgington (1995) both pointed out that inferring *if not- p then q* from p or q is invalid for the type of non-material conditional that Stalnaker specifies and for the probability conditional. Their views have influenced the account of this inference form given below (based on Gilio and Over, 2012; and Over et al., 2010). However, this account differs in a number of significant ways from their analyses. There was some research in the binary paradigm on this argument form, but it did not manipulate the justification or probability of p or q before *if not- p then q* was inferred (Ormerod and Richardson, 2003). As we have pointed out above, it is essential to study inferences with uncertain premises in the new paradigm, and this point also applies to the premise p or q , as can now be illustrated.

Consider a woman who confidently believes the following disjunction on general grounds to do with a strong belief in global warming:

- (4) The Thames Barrier will be reinforced (r) or London will be flooded (l).

Intuitively, she will also have solid grounds for inferring the following conditional from (4):

- (5) If the Thames Barrier is not reinforced, then London will be flooded.

This woman believes the disjunction (4) as a result of *non-constructive* justification. She does not infer (4) from 'below', from *r* alone or from *l* alone. She infers *r* or *l* from 'above', from general beliefs about global warming causing a rise in sea levels, and cannot say whether *r* holds or *l* holds at the time of the inference. Using the Ramsey test to assess (5), she is going to suppose *not-r* and make whatever changes are necessary to maintain consistency. She will remain consistent if she holds on to her beliefs about global warming and consequently her high confidence in (4). From her confidence in (4) plus the supposition of *not-r*, she can confidently infer *l*, giving her high confidence in (5) by the Ramsey test.

In contrast, consider a man who confidently believes the following disjunction simply on the grounds of a belief that his car has been stolen:

- (6) My car has been stolen (*c*) or destroyed in an accident (*d*).

Now intuitively, the man has weak grounds for inferring this conclusion:

- (7) If my car has not been stolen, then it has been destroyed in an accident.

The man believes the disjunction (6) as a result of a *constructive* justification. He does infer (6) from 'below', from *c* alone. Using the Ramsey test on (7), he is going to suppose *not-c* and make whatever changes are necessary to maintain consistency. He will remain consistent only if he suspends his belief in *c*, but that is his sole ground for believing in (6), and he cannot consistently retain his belief in (6) without any justification at all, at least as long as he has the general belief that his beliefs are justified. Without confidence in (6) in his hypothetical thought, he cannot infer *d* confidently, and he will have little confidence in (7) by the Ramsey test. What he might well infer is that, if his car has not been stolen, then it has been towed away for a parking violation. It might be asked why he would infer (6) from *c*, but there can be a number of reasons for inferring a disjunction from one of its disjuncts. For example, he might have an insurance policy stating that, if his car has been stolen or destroyed in an accident, then the replacement cost will be paid. From this conditional and (6), justified by *c*, he can infer that the insurance company will pay the replacement cost. (Notice that we also illustrate in this last example that we are using the standard inclusive *or* in our analyses: (6) is true when both *c* and *d* are true. The insurance company must pay up even if the car has been stolen and destroyed by the thieves in an accident.)

The distinction between constructive and non-constructive justification is a generalization of the distinction between constructive and non-constructive proof in logic and mathematics (Dummett, 2000; Franklin and Daoud, 1996). The strength of the inference from (4) to (5), relative to the inference from (6) to (7), can be explained by this distinction and the Ramsey test (Gilio and Over, 2012; Over et al., 2010). The difference between the (4) to (5) inference and the (6) to (7) inference cannot lie in the degree of uncertainty of the premises, (4) and (6), since (4) and (6) could be equally uncertain or (equivalently) probable. The (6) to (7) inference illustrates most clearly that inferring *if not-p then q* from *p or q* cannot be a valid inference. The premise (6) is probable, and the conclusion (7) is improbable, when it is probable that the car has been stolen, but it is improbable it has been destroyed in an accident given that it has not been stolen. A single premise inference is invalid if its premise can be

(coherently) more probable than its conclusion, as we have explained above. The inference was wrongly considered valid in the old binary paradigm (Johnson-Laird and Byrne, 1991: 64, 2002: 650) simply because that did not consider ways in which the premise could be more probable than the conclusion. A full analysis in the new paradigm demonstrates that the inference is logically invalid, but also when it is probabilistically strong, and when probabilistically weak, according to how it has been justified (Gilo and Over, 2012). And there is experimental evidence that people do not confidently endorse inferring *if not- p then q* from *p or q* when *p or q* is constructively justified (Over et al., 2010; see also Cruz et al. 2015).

5.4 Belief Revision and Updating

Yet another benefit of the new paradigm in psychology is that it can give an account of *diachronic and dynamic reasoning* as well as *synchronic reasoning*. The old paradigm was severely limited by its implicit presupposition that reasoning only takes place in one temporal context, synchronically, and from assumptions and not changing beliefs. It almost completely overlooked the central role of reasoning in *belief revision* and *updating* through time. Consider again the man who is highly confident in (6). According to the old paradigm, if he adds to (6) the premise that his car has not been stolen, he should infer that it has been destroyed in an accident. This inference, from *c or d* and *not- c* to *d* , is also valid, *p*-valid, in the new paradigm. But unlike the old paradigm, the new paradigm can go on to cover diachronic reasoning and its role in belief revision. Recall that this man is confident in (6) mainly because he believes that his car has been stolen. Supposing he has some rationality, or indeed common sense, he will lose confidence in (6) when he gets evidence that his car has not been stolen. As he becomes more and more confident over time that his car has not been stolen, he will become more and more uncertain of (6). He will never have high confidence in both *c or d* and *not- c* , which would allow him to infer relatively high confidence in *d* . He starts out having high confidence in *c or d* and low confidence in *not- c* , and then he revises his degrees of belief to have high confidence in *not- c* and low confidence in *c or d* . The high uncertainty in one of the premises always makes him highly uncertain of the conclusion, although of course the uncertainty of the conclusion for him should not, normatively, be greater than the sum of the uncertainties of the premises (see Oaksford and Chater, 2013, for much more on this inference form).

The Bayesian study of degrees of belief, expressed in subjective probability judgments, and inference has always been centrally concerned with belief updating. In its account of this updating, we begin with a *prior* degree of belief in a hypothesis *h* and likelihood judgments: degrees of belief in some evidence *e* given *h* and in *e* given *not- h* . We then use Bayes's theorem to infer a degree of belief in *h* given *e* (Howson and Urbach, 2005). When we later observe (perhaps after running an experiment) that *e* (and nothing stronger) definitely holds, we would use *strict conditionalization* to infer a *posterior* degree of belief in *h* , equal to the degree of belief in *h* given *e* . As explained above, the subjective probabilities of conditionals are seen in the new paradigm as subjective conditional probabilities: *$P(\text{if } h \text{ then } e)$* is *$P(e|h)$* , *$P(\text{if not-}h \text{ then } e)$* is *$P(e|not-h)$* , and *$P(\text{if } e \text{ then } h)$* is *$P(h|e)$* . With experimental support for these

identities in ordinary people's judgments (as in Evans et al., 2003; Over et al., 2007), a Bayesian account can be given of conditional belief updating using ordinary indicative conditionals (as in Oaksford and Chater, 2007, 2009, 2013).

To be more precise, degrees of beliefs at different times should be distinguished. Suppose at time one we begin with a prior degree of belief in h , $P_1(h)$, along with the relevant likelihood judgments $P_1(e|h)$ and $P_1(e|not-h)$. Suppose also that we conform to Bayes's theorem, inferring a degree of belief in h given e at time one, $P_1(h|e) = x$. We begin with $P_1(e) < 1$ at time one, but move on to $P_2(e) = 1$ at time two after an observation or experiment. Assuming the *rigidity* or *invariance* of the conditional probability $P_1(h|e) = P_2(h|e)$, we can finally use strict conditionalization to update our degree of belief in h by inferring diachronically $P_2(h) = P_1(h|e) = x$ from $P_2(e) = 1$.

To generalize the case in which the minor premise becomes more probable over time but remains uncertain, $P_1(e) < P_2(e) < 1$, *Jeffrey conditionalization*, or *Jeffrey's rule*, could be used, again assuming rigidity or invariance in the conditional probabilities:

$$P_2(q) = P_2(p)P_1(q|p) + P_2(not-p)P_1(q|not-p)$$

Oaksford and Chater refer to conditionalization, both strict and Jeffrey, and the assumption of rigidity, in their theory of conditional reasoning (Oaksford and Chater, 2007, 2009; Over, 2009). If no judgment can be made about the conditional probability of q given *not-p*, an interval can be given, as detailed above. For the new paradigm, the use of strict or Jeffrey conditionalization can be seen as an instance of *diachronic*, or *dynamic*, MP: inferring some (interval) confidence at time two in h from confidence in *if e then h* at time one and confidence in e at time two. Diachronic MP is p-valid as long as invariance in conditional probability holds over time, but sometimes learning extra information, in addition to e , implies a failure of invariance. Then the inference is no longer justified, and the conclusion h can be withdrawn, making this updating or revising reasoning *non-monotonic* (Oaksford and Chater, 2007, 2009, 2013; Over, 2009).

There is a normative debate about both types of conditionalization in the foundations of Bayesianism and the philosophy of science (Howson and Urbach, 2005; Jeffrey, 1991b), and the descriptive adequacy of Jeffrey conditionalization has also been questioned (Over and Hadjichristidis, 2009; Zhao and Osherson, 2010). And there is considerable debate, in the psychological field of judgment and decision making, about the extent to which people conform to Bayes's theorem (Barbey and Sloman, 2007). Nevertheless, giving an account of human conditionalization, diachronic reasoning, and belief updating, and the extent to which these inferences are fallacious or biased, must be a basic aim of the new paradigm of reasoning, which has the great advantage that it can be combined and integrated with research on probability judgment and decision making (see also Baratgin and Politzer, 2010, on belief updating). It could, for instance, describe how people's confidence in (1) might increase over time, as well as their confidence in the prediction that global warming will continue, and account for their inferred confidence that London will be flooded, and the decision making that would lead to in turn.

5.5 Conclusion

The old paradigm in the psychology of reasoning was much too quick and simplistic in trying to develop a theory of human reasoning. There was supposedly a deep debate in the old paradigm about whether this theory should be one of ‘mental logic’ or ‘mental models’. From the new point of view, however, these theories are only slightly different examples of the same old paradigm. In the mental logic theory of Braine and O’Brien (1991), a natural deduction system for binary extensional logic was modified, mainly by changing certain rules for the conditional. But this still failed to block all the paradoxes of the truth functional conditional and left the resulting system without an adequate semantics (Evans and Over, 2004). In the mental models theory of Johnson-Laird and Byrne (1991), what is essentially presented is a simplified semantic tableau rule system (like that of Jeffrey, 1994) for binary extensional logic, and the ordinary indicative conditional is treated as equivalent to the truth functional conditional, with all the problems that that brings with it (Evans and Over, 2004). Neither of these theories covered inference from uncertain beliefs nor accounted for belief revision and updating, and neither theory could be integrated with judgment and decision making. It is also possible to postulate modified mental rules for probability logic, since p-validity can be axiomatized (Adams, 1998), and a proposal of this type is at an early stage of development (Pfeifer, 2013; Pfeifer and Kleiter, 2010; Over and Cruz, 2018). There is much to do to formulate in detail an adequate theory of reasoning from uncertain beliefs, to test it further in experiments, and to integrate it fully with judgment and decision making, but at least these have become attainable goals in the new paradigm in the psychology of reasoning.

References

- Adams, E. (1975) *The Logic of Conditionals: An Application of Probability to Deductive Logic*. Dordrecht: Reidel.
- Adams, E. (1998) *A Primer of Probability Logic*. Stanford: CLSI Publications.
- Baratgin, J., Over, D. E., and Politzer, G. (2013). Uncertainty and de Finetti tables. *Thinking and Reasoning* 19: 308–28.
- Baratgin, J. and Politzer, G. (2010) Updating: A psychological basic situation of probability revision. *Thinking and Reasoning* 16: 245–87.
- Barbey, A. K. and Sloman, S. A. (2007) Base-rate respect: From ecological rationality to dual processes. *Behavioral and Brain Sciences* 30: 241–97.
- Braine, M. D. S. and O’Brien, D. P. (1991) A theory of If: A lexical entry, reasoning program, and pragmatic principles. *Psychological Review* 98: 182–203.
- Byrne, R. M. J. and Johnson-Laird, P. N. (2009) ‘If’ and the problems of conditional reasoning. *Trends in Cognitive Sciences* 13: 282–7.
- Byrne, R. M. J. and Johnson-Laird, P. N. (2010). Conditionals and possibilities. In M. Oaksford and N. Chapter (eds), *Cognition and Conditionals*, Oxford: Oxford University Press, pp. 55–68.
- Cruz, N., Baratgin, J., Oaksford, M., and Over, D. E. (2015) Bayesian reasoning with ifs and ands and ors. *Frontiers in Psychology* 6: 192.
- Cruz, N. and Oberauer, K. (2014) Comparing the meanings of ‘if’ and ‘all’. *Memory and Cognition* 42: 1345–56.

- de Finetti, B. (1936/1995) The logic of probability (translation of 1936 original). Translated in R. B. Angell, The logic of probability. *Philosophical Studies* 77: 181–90.
- de Finetti, B. (1937/1964) Foresight: Its logical laws, its subjective sources (translation of 1937 original). In H. E. Kyburg and H. E. Smokier (eds), *Studies in Subjective Probability*. New York: Wiley, pp. 55–118.
- Douven, I. (2008) Kaufmann on the probabilities of conditionals. *Journal of Philosophical Logic* 37: 259–66.
- Douven, I. (2011) Indicative conditionals. In L. Horsten and R. Pettigrew (eds), *A Companion to Philosophical Logic*. London: Continuum Press, pp. 383–405.
- Douven, I. (2016) *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.
- Douven, I., and Dietz, R. (2011) A puzzle about Stalnaker's hypothesis. *Topoi* 30: 31–7.
- Dummett, M. (2000) *Elements of Intuitionism*. Oxford: Clarendon.
- Edgington, D. (1995) On conditionals. *Mind* 104: 235–329.
- Edgington, D. (2003) What if? Questions about conditionals. *Mind and Language* 18: 380–401.
- Edgington, D. (2004) Counterfactuals and the benefit of hindsight. In P. Dowe and P. Noordhof (eds), *Cause and Chance: Causation in an Indeterministic World*, London: Routledge, pp. 12–27.
- Edgington, D. (2008) Counterfactuals. *Proceedings of the Aristotelian Society* 108: 1–21.
- Edgington, D. (2014) Evaluating conditional chances and evaluating counterfactuals. *Studia Logica* 102: 691–707.
- Elqayam, S. and Evans, J. St. B. T. (2011) Subtracting 'ought' from 'is': Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences* 34: 233–90.
- Elqayam, S. and Over, D. E. (2013) New paradigm psychology of reasoning: An introduction to the special issue edited by Elqayam, Bonnefon, and Over. *Thinking and Reasoning* 19: 249–65.
- Evans, J. St. B. T., Handley, S., Neilens, H., and Over, D. E. (2007) Thinking about conditionals: A study of individual differences. *Memory and Cognition* 35: 1772–84.
- Evans, J. St. B. T., Handley, S. J., and Over, D. E. (2003) Conditional and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 321–35.
- Evans, J. St. B. T., and Over, D. E. (2004) *If*. Oxford: Oxford University Press.
- Evans, J. St. B. T., and Over, D. E. (2009) Conditional truth: Comment on Byrne and Johnson-Laird. *Trends in Cognitive Sciences* 14: 5.
- Evans, J. St. B. T., Thompson, V., and Over, D. E. (2015) Uncertain deduction and conditional reasoning. *Frontiers in Psychology* 6: 398.
- Farrell, R. J. (1979) Implication and presupposition. *Notre Dame Journal of Formal Logic* 27: 51–61.
- Franklin, J. and Daoud, A. (1996) *Proof in Mathematics: An Introduction*. Sydney: Quakers Hill Press.
- Fugard, J. B., Pfeifer, N., Mayerhofer, B., and Kleiter, G. D. (2011) How people interpret conditionals: Shifts toward conditional event. *Journal of Experimental Psychology-Learning Memory and Cognition* 37: 635–48.
- Gauffroy, C. and Barrouillet, P. (2009) Heuristic and analytic processes in mental models for conditional: An integrative developmental theory. *Developmental Review* 29: 249–82.
- Gilio, A. (2002) Probabilistic reasoning under coherence in System P. *Annals of Mathematics and Artificial Intelligence* 34: 5–34.
- Gilio, A. and Over, D. E. (2012) The psychology of inferring conditionals from disjunctions: A probabilistic study. *Journal of Mathematical Psychology* 56: 118–31.
- Giroto, V. and Johnson-Laird, P. N. (2004) The probability of conditionals. *Psychologica*, 47: 207–25.

- Hahn, U. and Oaksford, M. (2007) The rationality of informal argumentation: A Bayesian approach to reasoning fallacies. *Psychological Review* 114: 704–32.
- Howson, C. and Urbach, P. (2005) *Scientific Reasoning: The Bayesian Approach* (Third Edition). Chicago: Open Court.
- Jeffrey, R. C. (1991a) Matter of fact conditionals. *Aristotelian Society Supplementary Volume* 65: 161–83.
- Jeffrey, R. C. (1991b) *The Logic of Decision* (Second Edition). Chicago: Chicago University Press.
- Jeffrey, R. C. (1994) *Formal Logic: Its Scope and Limits* (Third Edition). New York: McGraw Hill.
- Johnson-Laird, P. N. and Byrne, R. M. J. (1991) *Deduction*. Hove & Hammersmith, UK: Erlbaum.
- Johnson-Laird, P. N. and Byrne, R. M. J. (2002) Conditionals: A theory of meaning, pragmatics, and inference. *Psychological Review* 109: 646–78.
- Kaufmann, S. (2004) Conditioning against the grain. *Journal of Philosophical Logic* 33: 583–606.
- Lewis, D. K. (1976) Probabilities of conditionals and conditional probabilities. *Philosophical Review* 85: 297–315.
- Manktelow, K. I., Over, D. E., and Elqayam, S. (2011). *Paradigm Shift: Jonathan Evans and the Science of Reason*. In K. I. Manktelow, D. E. Over, and S. Elqayam (eds), *The Science of Reason: A Festschrift for Jonathan St B T Evans* (1–16). Hove, UK: Psychology Press.
- Milne, P. (2012) Indicative conditionals, conditional probabilities, and the ‘defective truth-table’: A request for more experiments, *Thinking and Reasoning* 18: 196–224.
- Oaksford, M. and Chater, N. (2007) *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.
- Oaksford, M. and Chater, N. (2009) Précis of Bayesian rationality: The probabilistic approach to human reasoning. *Behavioral and Brain Sciences* 32: 69–84.
- Oaksford M. and Chater, N. (2013) Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking and Reasoning* 19: 346–79.
- Oberauer, K. and Wilhelm, O. (2003) The meaning(s) of conditionals: Conditional probabilities, mental models and personal utilities. *Journal of Experimental Psychology: Learning Memory and Cognition* 29: 680–93.
- Ormerod, T. C. and Richardson, H. (2003) On the generation and evaluation of inferences from single premises. *Memory and Cognition* 31: 467–78.
- Over, D. E. (2009) New paradigm psychology of reasoning. *Thinking and Reasoning* 15: 431–8.
- Over, D. E. (2017) Causation and the probability of causal conditionals. In M. Waldmann (ed.), *The Oxford Handbook of Causal Reasoning*, Oxford: Oxford University Press, pp. 307–25.
- Over, D. E. and Cruz, N. (2018) Probabilistic accounts of conditional reasoning. In L. J. Ball and V. A. Thompson (eds), *International Handbook of Thinking and Reasoning*, Hove, UK: Psychology Press, pp. 434–50.
- Over, D. E., Douven, I., and Verbrugge, S. (2013) Scope ambiguities and conditionals. *Thinking and Reasoning* 19: 284–307.
- Over, D. E., Evans, J. St. B. T., and Elqayam, S. (2010) Conditionals and non-constructive reasoning. In M. Oaksford and N. Chater (eds), *Cognition and Conditionals: Probability and Logic in Human Thinking*, Oxford: Oxford University Press, pp. 135–51.
- Over, D. E. and Hadjichristidis, C. (2009) Uncertain premises and Jeffrey’s rule. *Behavioral and Brain Sciences* 32: 97–8.
- Over, D. E., Hadjichristidis, C., Evans, J. St. B. T., Handley, S. J., and Sloman, S. A. (2007) The probability of causal conditionals. *Cognitive Psychology* 54: 62–97.

- Pfeifer, N. (2013) The new psychology of reasoning: A mental probability logical perspective. *Thinking and Reasoning* 19: 329–45.
- Pfeifer, N. and Kleiter, G. D. (2010) The conditional in mental probability logic. In M. Oaksford and N. Chater (eds), *Cognition and Conditionals: Probability and Logic in Human Thinking*, Oxford: Oxford University Press, pp. 153–73.
- Politzer, G. (2005) Uncertainty and the suppression of inferences. *Thinking and Reasoning* 11: 5–33.
- Politzer, G., Over, D. E., and Baratgin, J. (2010) Betting on conditionals. *Thinking and Reasoning* 16: 172–97.
- Quine, W. V. O. (1952) *Methods of Logic*. London: Routledge & Kegan Paul.
- Ramsey, F. P. (1926/1990) Truth and probability (originally published 1926). In D. H. Mellor (ed.), *Philosophical Papers*, Cambridge: Cambridge University Press, pp. 52–94.
- Ramsey, F. P. (1929/1990) General propositions and causality (originally published 1929). In D. H. Mellor (ed.), *Philosophical Papers*, Cambridge: Cambridge University Press, pp. 145–63.
- Singmann, H., Klauer, K. C., and Over, D. E. (2014) New normative standards of conditional reasoning and the dual-source model. *Frontiers in Psychology*, 5: 316.
- Skovgaard-Olsen, N., Singmann, H., and Klauer, K. C. (2016) The relevance effect and conditionals. *Cognition* 150: 26–36.
- Stalnaker, R. (1968) A theory of conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, Oxford: Blackwell, pp. 98–112.
- Stalnaker, R. (1975) Indicative conditionals. *Philosophia*, 5: 269–86.
- Stevenson, R. J. and Over, D. E. (1995) Deduction from uncertain premises. *Quarterly Journal of Experimental Psychology* 48A: 613–43.
- Stevenson, R. J. and Over, D. E. (2001) Reasoning from uncertain premises: Effects of expertise and conversational context. *Thinking & Reasoning* 7: 367–90.
- Swartz, N. and Bradley, R. (1979) *Possible Worlds: An Introduction to Logic and Its Philosophy*. Oxford: Basil Blackwell. http://www.sfu.ca/~swartz/modal_fallacy.htm.
- Tversky, A. and Kahneman, D. (1983) Extensional vs intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* 90: 293–315.
- Zhao, J. and Osherson, D. (2010) Updating beliefs in light of uncertain evidence: Descriptive assessment of Jeffrey's rule. *Thinking and Reasoning* 16: 288–307.

6

Counterfactuals to the Rescue

Cleo Condoravdi

6.1 Introduction

Edgington (2008) argues that the stakes of philosophical theorizing about conditionals are so high because of the role conditionals play in thought and reasoning.* While advocating an analysis of counterfactuals in terms of conditional probability, she singles out a central problem that all theories of counterfactuals must address:

All theories of subjunctives—Goodman’s, Lewis’s, mine—share what is essentially the same problem, that of specifying what you hang on to and what you give up when you make a counterfactual supposition: suppose such-and-such had been the case; what do you hold constant? For Goodman this is the problem of cotenability, for Lewis it’s the problem of closeness, for me, it’s the problem of which context shift is appropriate, which probability distribution is the appropriate one, given that it is not the one that represents your present state of belief. (pp. 13–14)

This chapter uses grammatical evidence from the phenomenon of polarity reversal in counterfactuals to shed light on the role context and, in particular, pragmatic presuppositions play in determining what you hold constant and what you give up when interpreting counterfactuals. It makes the case that counterfactuals exhibiting polarity reversal settle that the consequent follows not only from the explicitly mentioned antecedent but also from each of a set of evoked alternative antecedents. In order for this to happen, opening up the possibility of the antecedent requires also opening up the possibility of the alternatives. This, in turn, implies that in making the counterfactual assumption, you not only give up propositions you take for granted that are inconsistent with it, but also certain contextually entailed propositions even though they are consistent with the counterfactual assumption. The close study of grammatical systems shows that language endows simple words with the kind of meaning to enable compact reasoning about alternative possibilities without requiring of speakers to spell it all out.

* I would like to thank Lee Walters and Sven Lauer for discussion and comments on the chapter.

6.2 Polarity Reversal in Counterfactuals

Positive polarity items (PPIs) are expressions which are, generally, not acceptable within the scope of negation in the same clause. They are a rather heterogeneous lot, including the temporal adverbials *already* and *still*, as well as degree modifiers such as *pretty* and *rather*. (1)–(3) exemplify the regular distribution of three of these positive polarity items.¹

- (1) a. He has already arrived. / He has arrived already.
b. ?# He has not already arrived. / ?# He has not arrived already.
NOT > ALREADY
- (2) a. He is still at home. / He is at home still.
b. ?# He is not still at home. / He is not at home still.
NOT > STILL
- (3) a. She is pretty smart.
b. ?# She is not pretty smart.

The temporal adverbials can syntactically appear over or under negation, and the logical scope mirrors the overt syntactic scope. Specifically, if the negation appears to the left of the adverbial, as in (1-b), (2-b), it takes scope over it; if the negation appears to the right of the adverbial, as in (4), it scopes under the adverbial. Thus, (4-a) and (4-b) are on a par with (1-a) and (2-a), since the PPIs are not in the direct scope of negation.

- (4) a. He is already not at his post.
ALREADY > NOT
b. She is still not at her post.
STILL > NOT

Exceptionally, PPIs can appear in the scope of negation in certain environments. Baker (1970a,b), who first studied the phenomenon systematically, introduced the term ‘polarity reversal’ and identified the antecedent of counterfactual conditionals as an environment in which PPIs can appear in the scope of negation. (5) shows the exceptional acceptability of the three PPIs above within negated antecedents of counterfactuals.

- (5) a. If he had not already arrived, we would have postponed the meeting.
b. If he were not still at home, we would have missed him.
c. If she weren’t pretty smart, she wouldn’t have gotten out of this situation.

The counterfactual suppositions in (5)—*if he had not already arrived*, *if he were not still here*, *if she weren’t pretty smart*—all include a negation that scopes over a PPI. Therefore, (5-a), (5-b), and (5-c) should be at least as problematic as (1-b), (2-b), and (3-b), but they are not.

¹ As the variants in examples (1) and (2) illustrate, the temporal adverbials show a degree of freedom as to their syntactic placement.

A confounding factor that needs to be set aside is that (1-b), (2-b), and (3-b) appear acceptable in certain contexts, especially if *not* is emphatically stressed and the utterance is understood as an instance of denial. According to Horn (1989) and van der Sandt (1991), a denial is a speech act objecting to some aspect of the full informational content of a previous utterance, or even to the particular form of the utterance. Thus, it can reject that utterance for any number of disparate reasons: because the proposition it expresses is false, or because it suffers from presupposition failure, or because one of its implicatures is false, or because it gives rise to objectionable connotations or inferences of any kind. The polarity of a sentence used as a denial may be negative or positive.² Consequently, occurrences of (1-b), (2-b), or (3-b) can be acceptable if they engage with a previous utterance by a different speaker. By contrast, whenever a denial construal is disfavored, as in (6) and (7), PPIs resist being in the scope of negation.

- (6) a. John said he'd come late but he is already here.
b. # John said he'd come early but he is not already here.
- (7) a. John said he'd leave early but he is still here.
b. # John said he'd leave late but he is not still here.

The acceptability of (1-b), (2-b), and (3-b) under certain special conditions should, therefore, be set apart from cases of polarity reversal, which is systematic and, it would seem, grammatically conditioned. In polarity reversal, it is properties of the larger construction within which the negation and the PPIs appear which render the latter under the scope of the former licit.

As a general characterization of the environments that allow polarity reversal, Baker (1970a,b) proposed, roughly, that a sentence S_1 with a polarity sensitive expression exhibiting polarity reversal is acceptable if it entails or presupposes another sentence S_2 containing the same polarity sensitive expression whose polarity is the opposite of that of S_1 . The idea is that an S_1 containing a PPI within the direct scope of negation would be acceptable if it entails or presupposes some S_2 which contains the PPI in an affirmative context.

Baker (1970b) and, with certain amendments, Karttunen (1971) attributed polarity reversal in counterfactuals to the presupposition of the falsity of the antecedent, and regarded the phenomenon of polarity reversal in counterfactuals as striking evidence for the role of presupposition in accounting for polarity licensing. (8) is one of Baker's (1970b) examples, which, he claims, 'is not appropriate unless the speaker is willing to commit himself to the stronger presupposition in [(9)]' (p. 6).³

² For instance, in the mini-dialogues below, B's utterance, which objects to the truth-conditional content of A's utterance, has positive polarity. In (i) the polarity of the two sentences uttered by A and by B, respectively, is switched, whereas in (ii) it remains the same.

- (i) A: John never showed up.
B: What do you mean? John is already here.
- (ii) A: John left work early today.
B: What do you mean? John is still in his office.

³ Baker takes both *someone* and *already* to be positive polarity items. The status of *someone* as a PPI is controversial but that of *already*, which I will focus on here, is not. 'Stronger presupposition' is meant to compare the factual presupposition in (9) with a weaker non-factual presupposition.

- (8) If someone hadn't already succeeded in making radio contact with your husband, the Coast Guard would almost surely be expressing its concern.
- (9) Someone has already succeeded in making radio contact with your husband.

Where does the presupposition come from? Subjunctive conditionals⁴ do not necessarily presuppose that their antecedent is false (Anderson, 1951; Stalnaker, 1975; Edgington, 1995). In fact, they can be used to reason about either the truth or the falsity of the antecedent given what is known, or taken for granted, about the consequent. Both (10) and (11), for instance, are used in a context which is taken to be consistent with the antecedent. In (10) the conditional is used to argue for the truth of its antecedent given the truth of its consequent. In (11) the conditional is used to argue for the falsity of its antecedent given the falsity of its consequent, and, therefore, the falsity of the antecedent cannot be taken for granted in advance, i.e., it cannot be pragmatically presupposed prior to the utterance of the conditional.

- (10) If he had still been in his office then, the lights would have been on. The lights were on. Therefore, he was still in his office then.
- (11) If he had still been in his office then, the lights would have been on. The lights were off. Therefore, he was not in his office then.

Subjunctive conditionals with polarity reversal in the antecedent, however, can only have counterfactual uses. Baker observed that a subjunctive conditional with polarity reversal cannot be used to argue for the falsity of its antecedent, as evidenced by the contrast between (13) and (11). We can also observe that a subjunctive conditional with polarity reversal cannot be used to argue for the truth of its antecedent; consider the contrast between (12) and (10).

- (12) # If he hadn't still been in his office then, the lights would have been off. The lights were off. Therefore, he was not in his office then.
- (13) # If he hadn't still been in his office then, the lights would have been off. The lights were on. Therefore, he was still in his office then.

According to Baker, these contrasts are due to the fact that the PPI cannot appear in the scope of negation unless the conditional comes with the presupposition that its antecedent is false.

Although Baker's official principle governing when a PPI is acceptable under polarity reversal makes reference to a notion of presupposition as a relation between two sentences,⁵ in the discussion about the phenomenon, the notion of presupposition that he refers to is that of speaker presupposition.⁶ Karttunen (1971) argues that

⁴ I use the term 'subjunctive' to make reference to the form of a conditional, and the term 'counterfactual' for conditionals used in a context in which their antecedent is settled to be false. The term 'subjunctive' does not properly characterize the full range of morphosyntactic properties of the relevant conditionals, but I adopt it, as it is a standard term to pick out the particular class of conditionals which can have a counterfactual interpretation.

⁵ See the principle in (5b'') in Baker (1970b: 12).

⁶ This can be seen, for instance, in the quote given above, in connection with (8) and (9).

the polarity rule should not make reference to a relation between two sentences in isolation but rather be further relativized to a set of premises that are taken for granted. In the conclusion of his paper, Karttunen explicitly endorses the pragmatic notion of speaker presupposition.

[A]s sound as Baker's principle is, it is too narrow in requiring that there be some logical relation (entailment or presupposition) between the sentence which violates the general polarity rule and the corresponding sentence with reversed polarity. It is enough if the latter is regarded by the speaker as a necessary truth in the particular state of affairs that he is considering. One case where the speaker clearly should have such a belief is when the latter sentence stands in a certain logical relation to the former, that is, is either entailed or presupposed by it. It is this subset of polarity reversals that is explained by Baker's principle in its original form. Whatever the correct formulation ultimately turns out to be, it will have to cover a larger class of cases. (pp. 294–5)

The Baker–Karttunen generalization can be stated in a way that avoids assuming that subjunctive conditionals come with a presupposition regarding their antecedent: a PPI in the antecedent of a subjunctive conditional allows polarity reversal only if the conditional is counterfactual, i.e., it is uttered in a context which is incompatible with its antecedent.

In later work, Ladusaw (1980), Krifka (1991, 1995), Szabolcsi (2004), and Schwarz and Bhatt (2006) pursued alternative explanations for polarity reversal. Krifka and Szabolcsi attributed polarity reversal to the presence of a higher operator taking both the negation and the PPI in its scope; Ladusaw and Schwarz and Bhatt, to a special type of negation. In the terminology of Szabolcsi, the higher operator, or the special type of negation, 'rescues' the PPI. The aim then of these kinds of approaches is to characterize the types of rescuing operators. But with the exception of Schwarz and Bhatt, this work did not consider the case of counterfactuals.

Schwarz and Bhatt claim that rescuing involves a special negation, dubbed 'light negation'. They independently motivate this assumption based on certain distributional facts regarding the syntactic position of negation and other phrases within a clause in languages like German. They then show that the environments where negation has a special syntactic distribution largely coincide with the environments that support rescuing of PPIs.⁷ From a semantic point of view, on Schwarz and Bhatt's account, negation is not just two-way ambiguous between regular negation and light negation, but multiply ambiguous, as the interpretation of light negation has to vary across different environments in which it is supposed to occur. For counterfactuals, they state: 'We do not know for sure how light negation comes to enforce counterfactuality. But one possibility that comes to mind is that counterfactual light negation triggers a factive presupposition, that is, the presupposition that its scope is true' (p. 92). For instance, the negation in the antecedent of (8), which, given

⁷ Ladusaw (1980) also appeals to a special negation operator to explain the appearance of PPIs in the scope of negation. That operator, according to Ladusaw, has identical truth conditions with the standard negation operator but carries a conventional implicature that 'someone has believed until recently that the proposition in its scope is true'. Whereas Ladusaw intended his special negation operator to apply to the denial uses we saw in connection with (1-b), (2-b) and (3-b), Schwarz and Bhatt do not see light negation as covering cases of denial.

the PPI in its scope, has to be light negation, presupposes (9). This presupposition is then inherited by the entire conditional, as presuppositions originating in the antecedent do in general, and this is what imposes a counterfactual interpretation on the conditional.

Although it can derive the counterfactual interpretation of subjunctive conditionals in their role as PPI rescuers, the price this account pays is quite high given the multiple ambiguity of negation it has to assume. An analysis that does not postulate ambiguity of negation would seem to be preferable on general methodological grounds. Zeijlstra (2012) offers such an alternative account of Schwarz and Bhatt's distributional facts. But the issue of why counterfactuals act as PPI rescuers remains to be addressed.

A common feature of the approaches outlined above is that the status of an expression as a PPI does not follow from its lexical semantics, and for the particular case of polarity reversal in counterfactuals, that the interpretation of the counterfactual conditional itself does not play a role. In this chapter, I take as my starting point a more fine-grained view of polarity sensitivity, and trace the possibility of polarity reversal in counterfactuals to the meaning of the polarity sensitive expressions and the interpretation of counterfactuals. In my proposal, the negation in the antecedents of the conditionals in (5) is regular truth-functional negation, and the need for a counterfactual interpretation of subjunctive conditionals in their role as PPI rescuers will be derived. I will focus on the two temporal adverbials *already* and *still*, as their lexical semantics is better understood. Inevitably, in order for the analysis to cover other cases, the lexical semantics of the relevant PPIs would have to be settled on first. But the reasoning regarding the interpretation of counterfactuals should carry over.

The structure of the rest of the chapter is as follows. Section 6.3 gives more detail on the theory of PPIs that I am assuming. Section 6.4 fixes the meaning of *already* and *still*. Section 6.5 develops the analysis and shows how the rescuing behavior of making counterfactual assumptions comes about.

6.3 Scalar Assertions

The alternative analysis of PPI rescuing in antecedents of counterfactuals that I explore relies on the idea that, in virtue of their conventional meaning, polarity sensitive items are associated with a set of alternatives in addition to their ordinary truth-conditional content. These alternatives, ultimately, lead to a set of alternative propositions that are ordered by semantic strength with the proposition corresponding to the plain truth-conditional content of the sentence. The polarity of the sentence obviously figures crucially in the direction of the ordering by strength. The effect of the alternatives on the overall meaning is at the core of the informativity-based theories of polarity sensitivity proposed by Krifka (1995) and Chierchia (2006).

The limited distribution of one class of polarity sensitive expressions, including the PPIs of interest here, has to do with the fact that they give rise to *scalar assertions*. Scalar assertions are assertions of sentences that, semantically or pragmatically, evoke alternatives and in which the proposition based on the plain truth-conditional content is informationally ordered with respect to the alternatives. A scalar assertion

conveys more information than a plain assertion, as it negates any alternative propositions that are informationally stronger than the proposition based on the plain truth-conditional content.⁸

Plain and scalar assertions update the information state of a context, corresponding to what is common ground between the relevant agents in the context, as is familiar from Stalnaker (1973, 1975). If we construe information states as sets of possible worlds, the relation of strength (informativity) between such information states is in terms of the subset relation. With a slight abuse of notation but for ease of exposition, in the following, I use the same symbol for contexts and their information state parameter (the context set, in Stalnaker's terms). The basic effect of a plain assertion of ϕ in a context c can be identified with the update function $+$ defined as in (14). It relies on the notion of felicity, a necessary condition of which is presupposition satisfaction.⁹

- (14) $c + \phi = \{w \in c \mid w \in \llbracket \phi \rrbracket_c\}$ provided ϕ is felicitous relative to c , else undefined.
- (15) ϕ is felicitous relative to context c only if c satisfies ϕ 's presuppositions.
- (16) Context c satisfies presupposition p of ϕ iff c entails p .

The relation of informational strength between two sentences can then be defined in terms of the update function as in (17).

- (17) ϕ_1 is informationally at least as strong as ϕ_2 iff for any context c relative to which ϕ_1 and ϕ_2 are felicitous, $c + \phi_1 \subseteq c + \phi_2$.

Scalar assertions of a sentence with alternatives update the context with the proposition corresponding to the plain truth-conditional content and, in addition, negate the truth of any informationally stronger alternative propositions. Thus, in general, scalar assertions result in a more informative context than plain assertions. The effect of a scalar assertion on a context is defined in (18).

- (18) $ScalAssert(\langle \phi, Alt(\phi) \rangle, c) =$
 $\{w \in c \mid w \in \llbracket \phi \rrbracket_c \wedge \neg(\exists \phi' \in Alt(\phi) (w \in \llbracket \phi' \rrbracket_c \wedge c + \phi' \subset c + \phi))\}$

Polarity sensitive expressions with semantically triggered alternatives of a certain kind give rise to scalar assertions. Their acceptability depends on the relation of informational strength between the plain content and the alternatives. They are acceptable in an environment when the plain truth-conditional content, which, of course, depends on the polarity of that environment, is at least as strong as each one of the alternatives. In that case, $c + \phi = ScalAssert(\langle \phi, Alt(\phi) \rangle, c)$. By contrast, in an environment in which they are not acceptable, the plain truth-conditional

⁸ Utterances of *John ate some of the cookies*, for instance, constitute scalar assertions whenever *all*, the lexical alternative to *some*, is pragmatically evoked. A scalar assertion of this sentence is more informative than the assertion of its truth-conditional content because it also conveys the information corresponding to the scalar implicature that John did not eat all of the cookies.

⁹ The requirement of presupposition satisfaction for successful contextual update thus connects linguistically triggered presuppositions with pragmatic presuppositions.

content is informationally weaker than any one of the alternatives. In that case, $\text{ScalAssert}(\langle\phi, \text{Alt}(\phi)\rangle, c) = \emptyset$ even when $c + \phi \neq \emptyset$.

On this kind of theory, the contrast between (1-a) and (1-b), or that between (2-a) and (2-b), would be explained by the fact that the contribution *already* and *still* make to the meaning of the sentences generally results in consistent scalar assertions of (1-a) and (2-a) in contexts relative to which they are felicitous, whereas (1-b) and (2-b) lead to inconsistency in any such context. I show this in the following section.

If we could show that (5-a) and (5-b) are like (1-a) and (2-a) in this respect, rather than (1-b) and (2-b), the rescuing behavior of counterfactuals would follow. This is what I do in Section 6.5, where I show that with polarity reversal the plain content of the conditional is informationally stronger than each one of the alternatives, at least when the conditional is counterfactual.

6.4 *Already, Still*: Content, Presupposition and Alternatives

The temporal adverbials *already* and *still* have a trivial truth-conditional content but give rise to non-trivial implications. For example, (19-a) and (20-a) have the same truth-conditional content, shared with (21). In contrast to (21), they have an implication regarding his being in the cave prior to the reference time, in this case the time of utterance. The polarity of this implication is flipped for the two adverbials, as seen in (19-b) vs. (20-b). They also have the counter-to-expectation implication in (19-c) and (20-c).

- (19) a. He is already in the cave.
b. He was not in the cave earlier.
c. He got into the cave earlier than expected.
- (20) a. He is still in the cave.
b. He was in the cave earlier.
c. He has stayed in the cave longer than expected.
- (21) He is in the cave (now).

The first kind of implication is actually one aspect of a more general implication about a potential change. This implication has the hallmarks of a linguistically triggered presupposition, as it projects through possibility modals and antecedents of conditionals. For instance, (22-a) and (22-b) imply that the treasure was in the cave at an earlier time but may have been moved away, while (23-a) and (23-b) imply that the treasure was not in the cave at an earlier time but may have been moved there.

- (22) a. Maybe the treasure is still in the cave.
b. If the treasure is still in the cave, it will be safe.
- (23) a. Maybe the treasure is already in the cave.
b. If the treasure is already in the cave, it will be safe.

With predicates where the requisite presuppositions cannot be satisfied in normal contexts, use of the adverbials can lead to infelicity, as in (24-b) and (25-a).

- (24) a. It is already late.
b. # It is still late.
- (25) a. # It is already early.
b. It is still early.

Löbner (1989) characterized the meaning of these adverbials in terms of admissible intervals; admissible intervals consist of a positive/negative phase followed by a negative/positive phase but no other phases.¹⁰ As he put it, *already* 'pick[s] out a well-defined interval out of the overall time-axis, i.e., the time interval starting with the last negative phase beginning before [the reference time] and ending with the eventually following positive phase Starting from the fact that there is a positive phase of *p* which began before [the reference time], the question [for *still* is whether this phase continues until [the reference time] or has been succeeded by a negative phase' (p. 174).

Recasting Löbner's proposal somewhat, we can say that both adverbials presuppose that there is a transition within an interval with respect to a given property. The property is determined by the predicate the adverbials modify, in the case at hand, the clausal predicate. Let's call such intervals *I* periods of transition with respect to the property expressed by the clausal predicate. For *already*, the presupposed transition is from a maximal initial subinterval I_{NEG} of *I* not satisfying the property to a maximal final subinterval I_{POS} satisfying that property. For *still*, the presupposed transition is from a maximal initial subinterval I_{POS} of *I* satisfying the property to a maximal final subinterval I_{NEG} not satisfying that property. Therefore, *already* requires periods of transition such that $I_{\text{NEG}} < I_{\text{POS}}$, while *still* requires periods of transition such that $I_{\text{POS}} < I_{\text{NEG}}$. Both *already* and *still* assert of the reference time that it is within I_{POS} but because of their different presupposition their implications differ.

Let t_0 be the relevant reference time. Utterances of (22-a) and of (23-a) presuppose that t_0 is within periods of transition with respect to the temporal property of the treasure being in the cave. There could well be uncertainty in the common ground about what times on the time axis the periods *I*, I_{POS} , I_{NEG} , or even t_0 , correspond to (different times in different possible worlds), as well as the position of t_0 within any given period of transition. Utterances of (22-a) and of (23-a) convey that it is consistent with the relevant agent's information state that t_0 is within the positive phase I_{POS} of the period of transition, hence that the treasure may be in the cave at t_0 .

Already and *still* also activate alternatives. These are alternative times within the periods of transition at which the predicate they modify may hold. For *already* the alternative times follow the reference time, while for *still* they precede it.¹¹ Depending on when the transition occurs within a period of transition, the reference time may be within I_{NEG} , while at least some, if not all, of the alternative times are within I_{POS} . But if the reference time is within I_{POS} , then all the alternative times are also within I_{POS} .

¹⁰ Löbner (1989) studied the behavior of the German equivalents of *already* and *still*, *schon* and *noch*.

¹¹ The alternatives also give rise to the counter-to-expectation implication.

The association with alternatives along with their presuppositions is what makes *already* and *still* polarity sensitive.¹² Relative to a context c satisfying the presuppositions of $\text{Already}(\phi(t_0))$, the update with the plain truth-conditional content $c + \phi(t_0)$ is informationally stronger than the update with each one of the alternatives $c + \phi(t)$, where $t_0 < t$. Similarly for $\text{Still}(\phi(t_0))$, except that the alternatives are based on times $t < t_0$. Hence assertions of (1-a), (2-a), (19-a), (20-a), for instance, constitute scalar assertions. The relation of information strength is reversed when the polarity of the environment in which *already* and *still* is negative. Relative to a context c satisfying the presuppositions of $\neg \text{Already}(\phi(t_0))$,¹³ the update with the plain truth-conditional content $c + \neg\phi(t_0)$ is informationally weaker than the update with each one of the alternatives $c + \phi(t)$, where $t_0 < t$. Similarly for $\neg \text{Still}(\phi(t_0))$, where the alternatives are based on times $t < t_0$. Hence assertions of (1-b) and (2-b) constitute scalar assertions leading to an inconsistent update in any context in which their presuppositions are satisfied.

6.5 Counterfactuals and Alternatives

The truth-conditional, possible-world semantics for counterfactual conditionals of Stalnaker (1968) and Lewis (1973) relies on the notion of comparative similarity between worlds. For a counterfactual conditional to be true in a world w , the proposition expressed by the consequent of the conditional has to be true throughout the worlds in which the proposition expressed by the antecedent is true and which are otherwise as similar to w as can be.¹⁴ The facts of w determine the similarity to w , but not all facts have equal weight and the relation of similarity, as both authors have emphasized, is vague and context-dependent. Here is, for instance, how Lewis (1981) puts it:

We may think of factual background as ordering the possible worlds. Given the facts that obtain at a world i , and given the attitudes and understandings and features of context that make some of these facts count for more than others, we can say that some worlds fit the facts of i better than others do. Some worlds differ less from i , are closer to i , than others. (p. 218)

6.5.1 The semantics of counterfactuals and presuppositional antecedents

For purposes of formulating the semantics of conditionals, we can model maximal similarity in terms of selection functions which map a world w and a proposition p to a set of worlds in which p is true. For current purposes, we can assume that the value of a selection function is a non-empty set of worlds. Any selection function S also minimally satisfies the properties in (26) for any w and proposition p (construed as a set of worlds).¹⁵

¹² See also Krifka (1995).

¹³ Since presuppositions project through negation, the presuppositions of $\neg \text{Already}(\phi(t_0))$ are the same as those of $\text{Already}(\phi(t_0))$.

¹⁴ This informal statement presupposes what Lewis calls the Limit Assumption.

¹⁵ As Lewis (1973: 58–9) shows, these conditions characterize the selection functions ‘derived from (centered) systems of spheres satisfying the Limit Assumption’.

- (26) a. $S(w, p) \subseteq p$
 b. If $w \in p$, then $S(w, p) = \{w\}$
 c. If p entails q and $p \cap S(w, q) \neq \emptyset$, then $S(w, p) = \{w' \in S(w, q) \mid w' \in p\}$

On the traditional Stalnaker-Lewis semantics, the truth-conditional content of a subjunctive conditional *if* ϕ , *would* ψ , relative to a selection function S and a context c , is as in.¹⁶

- (27) $\llbracket \text{if } \phi, \text{ would } \psi \rrbracket_c^S = \{w \in W \mid S(w, \llbracket \phi \rrbracket_c) \subseteq \llbracket \psi \rrbracket_c\}$

The semantics in (27) may be adequate for conditionals whose antecedents bear no presuppositions, but it will not do for conditionals with presupposition-bearing antecedents. According to (27), the proposition that is given as an argument to the selection function corresponds to the (unrestricted) truth-conditional content of the antecedent. Below I show that the presuppositional content of the antecedent affects the content of a conditional and amend the semantics in (27) to incorporate the effect of presuppositions in the hypothetical assumption made by the antecedent.

As is well-known, the presuppositional implications of the antecedent of a conditional ‘project’ and become presuppositions of the entire conditional (Karttunen, 1974; Heim, 1992). Since the temporal adverbials *still* and *already* are associated with presuppositions, the antecedents of the conditionals in (28) and (29), as well as the conditionals themselves, carry the presuppositions triggered by the temporal adverbials.

- (28) If he were still in the cave, we would be getting worried.
 (29) If he were already in the cave, they would have to rush.

An utterance is felicitous only relative to contexts that satisfy (entail) its presuppositions. (28) and (29) are thus felicitous only in contexts which satisfy their respective presuppositions, entailing minimally, for (28), that he was in the cave earlier, and for (29), that he was not in the cave earlier.¹⁷

The presuppositional content of the antecedent not only imposes a felicity condition on the context in which the counterfactual is uttered, but it also affects the content of the counterfactual assumption.¹⁸ Compare (28) and (29) with (30) and (31).

¹⁶ I drop the parameter S for $\llbracket \phi \rrbracket$ and $\llbracket \psi \rrbracket$. If ϕ and ψ do not contain a counterfactual, S does not play a role in their interpretation.

¹⁷ As discussed in Section 6.4, the full presupposition requires that the reference time and the alternative times be within periods of transition with respect to the property of him being in the cave.

¹⁸ Heim (1992: 205), in her investigation of presupposition projection from antecedents of conditionals, makes the same observation about the effect of presuppositions on the content of counterfactuals: ‘Let me close this excursion with a remark on the effect of presuppositional requirements in the antecedent of a counterfactual’s truth conditions. Recall the context where Mary is presupposed to be in the phone booth. We noted above that an indicative *if*-clause like *If John is in the phone booth ...* in this context amounts to the supposition that both John and Mary are in the booth. This is otherwise for a minimally different subjunctive *if*-clause: If we say *If John WERE in the phone booth*, then it depends on the actual facts and the selection function whether the hypothetical situations under consideration have both people in the booth or have John there instead of Mary... *then Mary would be outside* is a felicitous and possibly true continuation. (As opposed to the deviant indicative variant *If John is in the phone booth, then Mary is*

(30) If he were in the cave (now), we would be getting worried.

(31) If he were in the cave (now), they would have to rush.

Although the truth-conditional content of the antecedents of all four conditionals is the same, the kind of counterfactual assumption (28) is making is different from that of (30), and, similarly, the one (29) is making is different from that of (31).¹⁹ The counterfactuals in (30) and (31) do not necessarily express the same content as the corresponding ones in (28) and (29), even if they are uttered in the same context (e.g., in a context in which it is taken for granted that he was earlier in the cave but no longer, or in a context in which it is taken for granted that he was not in the cave earlier, but that he would be eventually). For (30) or (31), the selection function may just pick worlds in which he is now in the cave, regardless of whether he was there earlier or not, or whether he would be there later on or not. By contrast, for (28) the selection function must pick worlds in which he was in the cave earlier and continues to be in the cave now, and similarly for (29) it must pick worlds in which he was not in the cave earlier but is in the cave now. So even if interpreted with respect to the same selection function S and context c , the interpretations of (28) and (30), or of (29) and (31), can differ. We can conclude that the proposition that is given as an argument to the relevant selection function is different for (28) and (30), and for (29) and (31).²⁰ If a context c satisfies (entails) the presuppositions of the antecedent ϕ of a counterfactual conditional, it encodes the necessary information, but since it entails $\neg\phi$, we cannot have it directly restrict the proposition that corresponds to the truth-conditional content of the antecedent (since $c \cap \llbracket \phi \rrbracket_c = \emptyset$).

To take into account the effect of presuppositions on the hypothetical assumption made by counterfactual conditionals, I use Heim's (1992) notion of a revised context and have the proposition expressed by the antecedent be contextually restricted by the revised context. $rev(c, \phi)$ as defined in (32) designates a particular kind of revision of c that allows for consistency with ϕ . It is the maximal (i.e., most inclusive and hence least informative) revision of c that still satisfies the presuppositions of ϕ .

(32) $rev(c, \phi) = \cup \{ X \subseteq W \mid c \subseteq X \text{ and } \phi \text{ is felicitous relative to } X \}$

outside. This is acceptable only if we are ready to conclude that Mary's being in the phone booth isn't presupposed after all.) This difference, of course, is predicted by [the context change potential of counterfactual conditionals]. But what is also predicted is that if we add to the subjunctive antecedent a *too*, as in *If John were in the phone booth too ...*, then the meaning is in a certain respect more like that of the indicative again: no matter what the selection function and facts of the world, we only get to consider hypothetical worlds with both people in the booth together. So *If John were in the phone booth TOO, then Mary would be outside* is also deviant.'

¹⁹ *Still* and *already* restrict the temporal reference of the antecedent to a (contextually determined) reference time, which obviously affects the antecedent's truth-conditional content. In order to control for that, I have made the temporal reference of the antecedents of (28)–(31) be uniformly the time of utterance (marked by the use of *were* as opposed to *had been*).

²⁰ This is not the only conceivable conclusion. An alternative would be to have presuppositions constrain the selection function chosen in a given context. Thanks to Lee Walters for discussion on this point.

If c entails $\neg\phi$, moving to $rev(c, \phi)$ opens up the possibility of ϕ . As Heim shows, $rev(c, \phi)$ satisfies the presuppositions of ϕ if and only if c does. If ϕ has no presuppositions, then $rev(c, \phi)$ would just be W .

We can now revise (27) to (33).²¹

$$(33) \quad \llbracket \text{if } \phi, \text{ would } \psi \rrbracket_c^S = \{w \in c \mid S(w, rev(c, \phi)) \cap \llbracket \phi \rrbracket_c \subseteq \llbracket \psi \rrbracket_c\}$$

The semantics in (33) is more radically context-dependent than the one in (27), as the proposition expressed by a conditional uttered in a context c is a subset of c 's context set, rather than a subset of the set of all possible worlds W .

Let us see more specifically what the revised context would be like for (28) and (29). Recall that the presupposition we have assigned to *still* and *already* is that the reference time and the alternative times are located within periods of transition with respect to the property of times expressed by the clausal predicate. Let us designate the property of times expressed by the clausal predicate in (28)–(31) as *he-in-cave*. While for the presuppositionless (30) and (31) uttered in a context c with reference time t_0 the proposition that is given as an argument to the relevant selection function S can be $\llbracket he-in-cave(t_0) \rrbracket_c$, it has to be a more restricted proposition for (28) and (29) so as to encode the presuppositional content.

(28) would be felicitous relative to context c only if c entails that the reference time t_0 and the alternative times, which are all times $t < t_0$, are located within a period of transition with respect to *he-in-cave*, with $I_{POS} < I_{NEG}$. In other words, c has to entail that he exits the cave at some point within the relevant period. Given that c entails that he is not in the cave at the reference time, then c is comprised solely of worlds in which the reference time t_0 is within I_{NEG} . For any alternative time t , c allows for possibilities in which t is within I_{NEG} and possibilities in which t is within I_{POS} . $rev(c, Still(he-in-cave(t_0)))$ opens up the possibility for *he-in-cave*(t_0), so it includes in addition worlds in which the reference time t_0 is within I_{POS} . Those will be worlds in which all the alternative times are within I_{POS} .

Similarly, (29) would be felicitous relative to context c only if c entails that the reference time t_0 and the alternative times, which are times $t_0 < t$, are located within a period of transition with respect to *he-in-cave*, with $I_{NEG} < I_{POS}$. In other words, c has to entail that he enters the cave at some point within the relevant period. c is comprised solely of worlds in which the reference time t_0 is within I_{NEG} . For any alternative time t , c allows for possibilities in which t is within I_{NEG} and possibilities in which t is within I_{POS} . $rev(c, Already(he-in-cave(t_0)))$ opens up the possibility for *he-in-cave*(t_0), so it includes in addition worlds in which the reference time t_0 and consequently all the alternative times are within I_{POS} .

The respective propositions that are given as arguments to the selection function are as in (34).

- (34) a. $rev(c, Still(he-in-cave(t_0))) \cap \llbracket he-in-cave(t_0) \rrbracket_c$
 b. $rev(c, Already(he-in-cave(t_0))) \cap \llbracket he-in-cave(t_0) \rrbracket_c$

²¹ To take into account presuppositions of the consequent the semantics would have to be further revised but I will ignore this here.

$rev(c, Still(he-in-cave(t_0)))$ preserves the information in c that the reference time t_0 and the alternative times are within a period of transition, and similarly for $rev(c, Already(he-in-cave(t_0)))$. What changes is the location of the transition between I_{POS} and I_{NEG} (for *still*), or between I_{NEG} and I_{POS} (for *already*), and, therefore, whether t_0 and the alternative times fall within the positive phase or the negative phase of the relevant period. For instance, for (28) the information that he exits the cave is preserved; what changes is the time of his exit from the cave, so as to allow for the reference time t_0 to be a time at which he is in the cave. For (29) the information that he enters the cave is preserved; what changes is the time of his entering the cave, so as to allow for the reference time t_0 to be a time at which he is in the cave.

For (30) and (31), by contrast, the propositions given as arguments to the selection function are less restrictive, so relative to the same world its values can be outside of $rev(c, Still(he-in-cave(t_0)))$ or of $rev(c, Already(he-in-cave(t_0)))$.

6.5.2 Antecedents with positive polarity

Let us first consider the role of the alternatives for counterfactuals which, like (28) and (29), have antecedents with positive polarity.²² Krifka (1995) and Chierchia (2006) propose that alternatives need not be exploited only at the level of the matrix sentence. This implies that within a complex expression the relative strength of the plain content vis-à-vis the alternatives can be assessed in an embedded position. In conditionals, specifically, this can happen at the level of the antecedent.

The set of alternatives for the antecedent of (28) is given in (35-a), and that for the antecedent of (29) in (35-b), where t_0 is the reference time.

- (35) a. $Alt(Still(he-in-cave(t_0))) = \{he-in-cave(t) \mid t < t_0\}$
 b. $Alt(Already(he-in-cave(t_0))) = \{he-in-cave(t) \mid t_0 < t\}$

As we saw in the previous section, the propositional argument of the selection function is a more restricted proposition than the proposition corresponding to the truth-conditional content of the antecedent. Therefore, in checking the relative strength of the plain content and the content of the alternatives we need to look at propositions restricted by the revised context. The sets of alternative propositions corresponding to (35) are as in (36).

- (36) a. $\{rev(c, Still(he-in-cave(t_0))) \cap \llbracket he-in-cave(t) \rrbracket_c \mid t < t_0\}$
 b. $\{rev(c, Already(he-in-cave(t_0))) \cap \llbracket he-in-cave(t) \rrbracket_c \mid t_0 < t\}$

None of the alternatives in (35-a) yields a more restricted proposition than (34-a), and similarly, none of the alternatives in (35-b) yields a more restricted proposition than (34-b). In fact, the proposition in (34-a) is stronger than each proposition in (36-a), and the proposition in (34-b) is stronger than each proposition in (36-b).²³ Therefore, making the counterfactual assumption that he is in the cave at t_0 is the

²² In the brief discussion in Section 6.4 on the polarity sensitivity of the adverbials, we considered contexts which satisfied their presuppositions and which were compatible with the plain content and the alternatives. When we move to antecedents of counterfactuals, the latter property, of course, no longer holds.

²³ On the reasonable assumption that the alternatives do not carry presuppositions of their own.

strongest relevant assumption that can be made.²⁴ Consequently, the PPIs *still* and *already* are acceptable in antecedents with positive polarity.

6.5.3 Antecedents with polarity reversal

When the antecedent of a subjunctive conditional exhibits polarity reversal, as in (5-a) and (5-b), repeated in (37), the relative assessment of strength between plain and alternative content can obviously not be done at the level of the antecedent, since negation reverses the strength relations.²⁵ Therefore, it has to be done at the level of the matrix sentence.

- (37) a. If he had not already arrived, we would have postponed the meeting.
b. If he were not still at home, we would have missed him.

We thus have to establish that assertions like (37-a) and (37-b) and, more generally, assertions of the schematic form in (38-a) and (38-b) constitute scalar assertions.

- (38) a. *if* $\neg$ (*Already*(ϕ)), *would* ψ
b. *if* $\neg$ (*Still*(ϕ)), *would* ψ

This means that the alternatives are themselves conditionals and we have to assess their relative strength vis-à-vis the plain content of the conditional uttered. The plain content of the conditionals in (38) is that of the conditional in (39-a) and their set of alternatives are as in (39-b).

- (39) a. *if* $\neg\phi$, *would* ψ
b. $\{ \text{if } \neg\phi', \text{ would } \psi \mid \phi' \in \text{Alt}(\phi) \}$

As discussed earlier, the content of both *Already*(ϕ) and *Still*(ϕ) is that the reference time t_0 is within the positive phase I_{pos} of the period of transition I determined by ϕ . The content of each alternative ϕ' based on alternative time t' is that t' is within the positive phase I_{pos} of I , where for *Already*(ϕ) every $t' \in \text{Alt}(t_0)$ is such that $t_0 < t'$, and for *Still*(ϕ) every $t' \in \text{Alt}(t_0)$ is such that $t' < t_0$.

Let us focus attention here on uses of (38-a) and (38-b) in contexts which are incompatible with the antecedent, that is contexts which entail ϕ .²⁶ In such cases the conditional corresponding to the plain content and the conditionals in the set of alternatives are all counterfactual, since ϕ contextually entails ϕ' for each $\phi' \in \text{Alt}(\phi)$.²⁷ For example, when (37-a) is interpreted counterfactually, the context entails that he had arrived by the reference time, and, therefore, that he had arrived by any time later than the reference time.

Let us abbreviate the revision of context c with $\neg$ (*Already*(ϕ)), or with $\neg$ (*Still*(ϕ)), as c_{rev} .²⁸ Given the semantics for subjunctive conditionals in (33), for the

²⁴ Given condition (26-c) on selection functions, this relation of strength gets reversed for the entire conditional.

²⁵ This mirrors the contrast between (1-a), (2-a) and (1-b), (2-b).

²⁶ In the next section we consider non-counterfactual interpretations of subjunctive conditionals and why they do not allow polarity reversal.

²⁷ Conversely, $\neg\phi'$ contextually entails $\neg\phi$, for any $\phi' \in \text{Alt}(\phi)$.

²⁸ Thus, depending on the particular case under consideration, c_{rev} is either $\text{rev}(c, \neg(\text{Already}(\phi)))$ or $\text{rev}(c, \neg(\text{Still}(\phi)))$.

plain content to be true in a world $w \in c$, it has to be the case that $S(w, c_rev \cap [\neg\phi]_c) \subseteq [\psi]_c$. Similarly, for any alternative to be true in w , it has to be the case that $S(w, c_rev \cap [\neg\phi']_c) \subseteq [\psi]_c$. Now, c_rev includes, for any $\phi' \in Alt(\phi)$, both possibilities in which $\neg\phi$ and ϕ' hold and possibilities in which $\neg\phi$ and $\neg\phi'$ hold. For instance, the c_rev corresponding to (37-a) includes worlds in which he has not arrived by the reference time t_0 but has arrived by some later time t , as well as worlds in which he has not arrived either by the reference time t_0 or by the later time t . The c_rev corresponding to (37-b) includes worlds in which he is not at home at the reference time t_0 but is at home at some earlier time t , as well as worlds in which he is at home neither at the reference time t_0 nor at the earlier time t . The fact that c_rev contains possibilities of both types guarantees that the propositions $c_rev \cap [\neg\phi']_c$ given as arguments to the selection function are not inconsistent. The question is, are the selection functions relative to which the conditionals are interpreted subject to any constraints such that the sets in (40) are related in a systematic way for every $w \in c$ and every $\phi' \in Alt(\phi)$?

- (40) a. $S(w, c_rev \cap [\neg\phi]_c)$
 b. $S(w, c_rev \cap [\neg\phi']_c)$

Given that the way the worlds are selected is based on the facts of w , the question could be formulated as follows: should the fact of a world $w \in c$ ²⁹ that some alternative ϕ' holds be preserved under the hypothetical assumption that $\neg\phi$? $c_rev \cap [\neg\phi]_c$ is consistent with the fact that ϕ' , since making the hypothetical assumption that the reference time is within the negative phase is compatible with an alternative time remaining within the positive phase of the period of transition determined by ϕ . If the worlds delivered by the selection function in (40-a) are exclusively $\neg\phi, \phi'$ worlds, then the sets in (40) would be disjoint and no connection could be established between the plain and the alternative content of counterfactuals with polarity reversal as in (38). But the fact that ϕ' , though consistent with the counterfactual assumption, may be given up and, in fact, may have to be given up.

Lewis (1979), in his explication of overall similarity, postulated that similarity of particular fact is of least or no importance. Analyses of counterfactuals based on premise semantics, such as Kratzer (1981, 1989) and Veltman (2005), have emphasized the role that dependence between facts plays in counterfactual reasoning. As Veltman (2005: 164) puts it: 'similarity of particular fact is important, but only for facts that do not depend on other facts. Facts stand and fall together. In making a counterfactual assumption, we are prepared to give up everything that depends on something that we must give up to maintain consistency.'

Now, for any given ϕ' and $w \in c$, the fact of w that ϕ and the fact that ϕ' are not independent from one another. Given what is presupposed, once ϕ is true throughout the worlds of c , so is any alternative ϕ' . These two types of facts are dependent, though not dependent due to a law-like generalization, or due to some causal connection, or due to a feature of the world that matters for overall similarity

²⁹ I am using 'fact of a world w ' in the familiar way to refer to a proposition true in w .

independently of the information in the context in which the conditional is used.³⁰ That ϕ' is true in w , for any $\phi' \in Alt(\phi)$, is a fact of w but it is true based on the fact that the transition from a positive to a negative phase (for *still*), or from a negative to a positive phase (for *already*), occurred when it did and the presupposition of a unique relevant transition within the period of transition. When we make the counterfactual assumption that the reference time is within the negative phase, this necessarily changes the time of the transition and it opens up the possibility that it could have occurred anywhere within the period of transition, including after the latest, or before the earliest, alternative time.

Therefore, the selection function can be assumed to be inclusive and not to distinguish between worlds in which the phase transition occurs after t_0 but before an alternative time t (these would be $\neg\phi, \phi'_t$ worlds) and worlds in which the phase transition occurs after t (these would be $\neg\phi, \neg\phi'_t$ worlds) in terms of their similarity to w . Given the constraint on selection functions in (26-c) and the fact that the proposition $c_rev \cap [\neg\phi']_c$ entails the proposition $c_rev \cap [\neg\phi]_c$, it follows that (41) holds.

$$(41) \quad S(w, c_rev \cap [\neg\phi']_c) \subseteq S(w, c_rev \cap [\neg\phi]_c)$$

With this property of the selection function, the plain content contextually entails each element in the set of alternatives, and therefore (38-a) and (38-b) constitute scalar assertions.

The more inclusive the value of the selection function, the less information it encodes about similarity of worlds. Polarity reversal is possible for counterfactuals uttered in a context with no information that would discriminate between those worlds in which the phase transition, if it hasn't occurred by t_0 , would occur earlier rather than later. That is arguably part of the presupposition carried by the polarity adverbials—not just the factual constraints on the context but also lack of conditional information about alternative timing of the transition within the period of transition, encoded via inclusive selection functions. Even if the internal structure of a world forces his being here sooner rather than later, or his staying here longer rather than shorter, use of the adverbials communicates and forces a kind of agnosticism about the time of the transition.

6.5.4 The role of counterfactuality

As discussed in Section 6.2, Baker observed that polarity reversal in subjunctive conditionals is accompanied by a counterfactual implication. Baker's generalization, reflected in his analysis, was that conditionals of the form in (38) are acceptable *if and only if* the conditionals presuppose *Already*(ϕ) and *Still*(ϕ), respectively.³¹ Schwartz and Bhatt took that generalization for granted and postulated the existence of a special negation that presupposes its prejacent, in the case of the conditionals of the

³⁰ This is arguably structurally analogous to Tichý's (1976) 'man with a hat' case, except for the different source of the dependence.

³¹ Since Baker relied on a semantic notion of presupposition (see Baker, 1970b: 15, n. 2), he must have assumed that subjunctive conditionals are ambiguous between those that presuppose the falsity of their antecedent and those that do not.

form in (38), *Already*(ϕ) and *Still*(ϕ), respectively.³² This presupposition originating in the antecedent projects to become a presupposition of the conditional. So conditionals of the type in (38) are predicted by both analyses to be acceptable only if uttered in contexts that entail *Already*(ϕ) or *Still*(ϕ), that is only in contexts satisfying the presuppositions of *Already* or *Still* and entailing ϕ .

But this turns out to be the wrong generalization. A conditional with polarity reversal in the antecedent can be counterfactual without the prejacent of the negation being entailed by the context. A case in point are conditionals with conjoined antecedents where the prejacent of the negation is not entailed by the context in which the conditional is used. PPIs in the scope of negation are acceptable as long as the other conjunct renders the conditional counterfactual. (42-a) is an instance where the negated part of the antecedent is *consistent* with the context in which the counterfactual is uttered, given the information conveyed by the first sentence. (42-b) is an instance where the negated conjunct is *entailed* by the context in which the counterfactual is uttered, given that the first sentence conveys (possibly even entails) that John is not here.³³

- (42) a. I am not sure if John has shown up but, luckily, we can do without him.
If, however, he had said he'd come help us and were not already here, I'd be annoyed with him.
- b. I don't mind that John hasn't shown up. If, however, he had said he'd come early and were not already here, I'd be annoyed with him.

The implication accompanying (42-a) is that John did not say that he'd come help us and that accompanying (42-b) that John did not say that he'd come early. So the conditionals are counterfactual, albeit not because of the negated conjunct containing the PPI.

The correct generalization then is that a PPI under the scope of negation is acceptable in the antecedent of a subjunctive conditional as long as the entire antecedent is counterfactual. In the case of conjunction it suffices for one of the two conjuncts to be incompatible with the context. In the examples in (42) the first conjunct's being counterfactual suffices for the full antecedent to be incompatible with the context of utterance and hence for the PPI in the scope of negation in the second conjunct to be 'rescued'. This suggests that the 'rescuing' of PPIs within negated antecedents of counterfactuals does not so much require that the prejacent of negation is presupposed to be true, but rather that the conditional itself is interpreted

³² The conditionals have the logical form in (i), where $\neg_L$ is the special light negation.

- (i) a. if $\neg_L$ (*Already*(ϕ)), would ψ
b. if $\neg_L$ (*Still*(ϕ)), would ψ

³³ In this respect, the examples in (42) are different from the cases of conjoined antecedents considered by Baker (1970b) and Karttunen (1971). In their examples, the negated conjunct with the PPI carried a presupposition triggered by an expression other than the PPI and satisfied by the first conjunct, itself incompatible with the context, e.g., *If John had a brother and his brother had not come to his help already*, where *his brother* presupposes that John had a brother. Unlike (42), where the negated conjunct is consistent with or even entailed by the context, in Baker's and Karttunen's examples the negated conjunct is simply infelicitous relative to the context.

against a context such that the relevant possibilities in which the antecedent is true—what the selection function delivers—are wholly outside the context.

The hypothetical assumptions made in (42) open up the issue of the timing of the transition and allow for possibilities in which the transition occurs anywhere within the relevant period of transition. As with the standard cases discussed in Section 6.5.3, for any world w in the context, the selection function does not distinguish between worlds in which the first conjunct is true and the transition occurs after the reference time t_0 but before an alternative time t and worlds in which the first conjunct is true and the transition occurs after both t_0 and t , regardless of the particular facts of the timing of the transition in w . Therefore, the selection function satisfies (41) and the conditionals in (42) can constitute scalar assertions.

As far as the question of why a subjunctive conditional needs to be counterfactual in order to be able to rescue PPIs, in the analysis pursued here, the answer will be found in how asserting the conditional in a context which is compatible with its antecedent influences the relation between the plain content of the subjunctive conditional and the content of the alternatives. Below I sketch out an argument that this is what happens in the two types of cases discussed in Section 6.2.

First let us see how the form of a conditional affects its interpretation. Stalnaker (1975) assumes that the information of a context in which a conditional is asserted restricts the selection functions used in its interpretation, stating:³⁴

[I]f the conditional is being evaluated at a world in the context set, then the world selected must, if possible, be within the context set as well... In other words, all worlds within the context set are closer to each other than any worlds outside it. The idea is that when a speaker says If A, then everything he is presupposing to hold in the actual situation is presupposed to hold in the hypothetical situation in which A is true. (pp. 275–6)

This amounts to a pragmatic constraint that all else being equal, any conditional χ asserted in context c has to be interpreted relative to selection functions S satisfying (43).

- (43) For any $w \in c$, $S(w, p) \subseteq c$, where p is the proposition determined by the antecedent of χ and by c as per the semantics of χ .³⁵

Stalnaker (1975), moreover, took the subjunctive marking in English conditionals to be a conventional means for signaling ‘that the selection function is one that may reach outside of the context set’ (p. 276).³⁶ This means that if a speaker wants to signal that (43) is not in force in the intended interpretation of his assertion of a conditional, he has to use the subjunctive form of the conditional, *if ϕ , would ψ* .

³⁴ Unlike (27) and (33), on Stalnaker’s semantics, selection functions pick a unique world rather than a set of worlds. Given the semantics in (33), a conditional asserted in a given context can only be true in worlds within its context set.

³⁵ As we have seen already, the context not only helps determine the truth-conditional content of the antecedent of a conditional but also the proposition that is given as an argument to the selection function. As specified in (33), for subjunctive conditionals with antecedent ϕ , the proposition that is given as an argument to the selection function is $rev(c, \phi) \cap \llbracket \phi \rrbracket_c$. As we will see in Section 6.5.5, for indicative conditionals it is $c \cap \llbracket \phi \rrbracket_c$.

³⁶ See also von Stechow (1998).

When ϕ is inconsistent with c , any selection function that satisfies (26-a) cannot simultaneously satisfy (43). In that case, therefore, any selection function S would be such that, for any $w \in c$, its value is fully outside the context set, i.e., $S(w, \text{rev}(c, \phi) \cap \llbracket \phi \rrbracket_c) \subseteq \text{rev}(c, \phi) \setminus c$. When ϕ is consistent with c , the worlds in c can vary in whether they verify ϕ . In those worlds $w \in c$ in which ϕ is true, the selection function would return $\{w\}$ as its value, given the centering condition in (26-b), and the conditional would be true only if ψ is true in w as well. So it is among the non- ϕ -worlds within c that the selection function can reach outside of c .

(44) and (45) exemplify the two types of non-counterfactual uses of subjunctive conditionals and the corresponding contrasts with antecedents with polarity reversal (repeated from (10), (12) and (11), (13) in Section 6.2).

- (44) a. If he had (still) been in his office then, the lights would have been on. The lights were indeed on. Therefore, he was (still) in his office then.
- b. # If he hadn't still been in his office then, the lights would have been off. The lights were indeed off. Therefore, he was not in his office then.
- (45) a. If he had (still) been in his office then, the lights would have been on. The lights were off. Therefore, he was not in his office then.
- b. # If he hadn't still been in his office then, the lights would have been off. The lights were on. Therefore, he was (still) in his office then.

Let us first consider what happens in the admissible cases in (44-a) and (45-a). In the first case the conditional is used to argue for the truth of the antecedent after the truth of the consequent is established. In the second case, the conditional is used to argue for the falsity of the antecedent after the falsity of the consequent is established. In both cases, we can assume that the initial context is compatible with both the antecedent ϕ and the consequent ψ of the conditional. In both cases, only those ϕ -worlds in the original context which are also ψ -worlds survive the update with the conditional.³⁷ In the first case, when the second premise comes in, any non- ψ -worlds (which would have to be non- ϕ -worlds) are removed. Finally, the conclusion, by stating that all remaining worlds are ϕ -worlds, implies that all the non- ϕ -worlds in the original context were non- ψ -worlds. In the second case, when the second premise comes in, all ψ -worlds (some of which could be ϕ -worlds) are removed. Finally, the conclusion, by stating that all remaining worlds are non- ϕ -worlds, implies that there were non- ϕ -non- ψ -worlds in the original context.³⁸

In order to explain why polarity reversal is not possible in subjunctive conditionals that are not counterfactual, we have to show that the plain content, *if* $\neg\phi$, *would* ψ , does not necessarily contextually entail *if* $\neg\phi'$, *would* ψ for every $\phi' \in \text{Alt}(\phi)$. Let us abbreviate the proposition $\text{rev}(c, \chi) \cap \llbracket \chi \rrbracket_c$ as $\mathcal{A}(c, \chi)$. Consider a context c which does not settle when the transition happened³⁹ and with some $w \in c$ such that $w \in \llbracket \neg\phi \rrbracket_c$, $w \in \llbracket \psi \rrbracket_c$ and $w \notin \llbracket \neg\phi' \rrbracket_c$ for some $\phi' \in \text{Alt}(\phi)$. This means that in w

³⁷ If there are no ϕ -worlds in the original context, the conditional is false at all the ϕ -worlds in that context, which would defeat the purpose of the first type of argument.

³⁸ Both types of argument thus establish a non-accidental connection between ϕ and ψ and between $\neg\phi$ and $\neg\psi$, e.g., in the examples above between his being in/away from the office and the lights being on/off.

³⁹ So the $\neg\phi$ -worlds in c are comprised of ϕ' -worlds and $\neg\phi'$ -worlds, for every $\phi' \in \text{Alt}(\phi)$.

one of the alternative times is within the positive phase of the period of transition determined by ϕ . For such worlds w , $S(w, \mathcal{A}(c, \neg\phi)) = \{w\}$ but $w \notin S(w, \mathcal{A}(c, \neg\phi'))$. So not only do the two sets not stand in the subset relation, they are disjoint. This disrupts the informational ordering between the plain semantic value of the conditional and the semantic value of at least one of its alternatives. The plain content is true in w but the value of $S(w, \mathcal{A}(c, \neg\phi'))$ may well be constituted of worlds which are not uniformly ψ -worlds. For instance, to the extent that the conditional is informative relative to c due to there being ϕ -worlds w' in c in which the conditional is false—which means that $S(w', \mathcal{A}(c, \neg\phi'))$ is not comprised of worlds which are uniformly ψ -worlds—the same could be the case of the alternative conditional *if* $\neg\phi'$, *would* ψ at worlds like w .

6.5.5 Indicative conditionals

If polarity reversal required that the preajcent of the negation be entailed by the context, as it does on Baker's and Schwarz and Bhatt's proposals, then polarity reversal would be completely excluded from indicative conditionals, which must be used in contexts compatible with their antecedent. No context can simultaneously satisfy both conditions, and hence it would be predicted that polarity reversal cannot be observed in the antecedents of indicative conditionals. However, polarity reversal can be observed with indicative conditionals provided the context is of a certain kind. The indicative conditionals with polarity reversal in (46-a) and (47-a) are acceptable and give rise to a kind of uniformity implication, spelled out in (46-b) and (47-b), respectively.

- (46) a. If he was not already in his office when they got there, they left without him.
- b. If he was not in his office when they got there, they left without him, regardless of when he came to the office.
- (47) a. If he was not still in town during their meeting, they did not see him.
- b. If he was not in town during their meeting, they did not see him, regardless of when he left town.

The reason polarity reversal is not *generally* acceptable in indicative conditionals is structurally equivalent to the reason polarity reversal is not acceptable with subjunctive conditionals when their antecedent is consistent with the context. Before showing this, we need to fix the semantics of the indicative conditional. Following Stalnaker's (1975) uniform treatment of subjunctive and indicative conditionals, I will assume that an indicative conditional *if* ϕ , ψ uttered in context c has the semantics in (48).

$$(48) \quad \llbracket \text{if } \phi, \psi \rrbracket_c^S = \{w \in c \mid S(w, c \cap \llbracket \phi \rrbracket_c) \subseteq \llbracket \psi \rrbracket_c\}$$

As with the semantics proposed for subjunctive conditionals in (33), the proposition expressed by an indicative conditional is a subset of the context's context set and the proposition that is given as an argument to the selection function encodes the presuppositions of the antecedent, in addition to its truth-conditional content.

The constraint in (43) amounts to (49) for an indicative conditional. (49) constrains the selection functions involved in (48) and plays a role in determining the plain and alternative content of conditionals like (46-a) and (47-a).

- (49) For any $w \in c$, $S(w, c \cap \llbracket \phi \rrbracket_c) \subseteq c$.

To see why polarity reversal is generally disallowed in indicative conditionals, we need to show that the relation of informational ordering between the plain content and the content of the alternatives is broken under polarity reversal. Adapting our notation from the previous section, let us now use $\mathcal{A}(c, \chi)$ to abbreviate the proposition $c \cap \llbracket \chi \rrbracket_c$. For those $w \in c$ in which χ is true, $S(w, \mathcal{A}(c, \chi)) = \{w\}$, given the centering condition in (26-b). Suppose the antecedent exhibits polarity reversal, as in (46-a) and (47-a). Consider a context c which does not settle when the transition happens and with some $w \in c$ such that $w \in \llbracket \neg\phi \rrbracket_c$, $w \in \llbracket \psi \rrbracket_c$, $w \notin \llbracket \neg\phi' \rrbracket_c$ for some $\phi' \in \text{Alt}(\phi)$. Then $S(w, \mathcal{A}(c, \neg\phi)) = \{w\}$ but $w \notin S(w, \mathcal{A}(c, \neg\phi'))$. As before, this breaks the subset relation between the values of S for the plain value of the antecedent and one of its alternatives, which in turn disrupts the informational ordering between the plain semantic value for the conditional and its alternative semantic values. To the extent that the conditional is informative relative to c , say due to there being a $\neg\phi \neg\psi$ -world in c , such a world may well be in $S(w, \mathcal{A}(c, \neg\phi'))$, which, by (49), is a subset of c .

Now suppose we have a context c favoring selection functions S satisfying the condition in (50).

- (50) For any $w \in c$ and any $\phi' \in \text{Alt}(\phi)$, $S(w, \mathcal{A}(c, \neg\phi')) \subseteq S(w, \mathcal{A}(c, \neg\phi))$

This ensures that any antecedent worlds ($\neg\phi$ -worlds) in the context behave uniformly across the alternatives. This means that the conditional supports limited strengthening of the antecedent inferences: when *if ϕ , ψ* is true, the conditionals with strengthened antecedents *if ϕ and ϕ' , ψ* and *if ϕ and $\neg\phi'$, ψ* are also true, for any $\phi' \in \text{Alt}(\phi)$. For instance, when (46-a) is true, so is (51), which is an instantiation of the uniformity implication of (46-a).

- (51) If he was not in his office when they got there and only came later, they left without him.

In such contexts, use of (46-a) or (47-a) would constitute a scalar assertion. For the counterexample we constructed above, even though $w \notin S(w, \mathcal{A}(c, \neg\phi'))$, any world in $S(w, \mathcal{A}(c, \neg\phi'))$ would be a ψ world, just like w is.

Conditionals like (46-a) and (47-a) convey that there is a bound on how late or how early the transition can occur and still have the consequent be true. Those features of a context that support a use of (46-a) as a scalar assertion, including the kinds of selection functions relevant for the interpretation of (46-a), would also verify the conditional in (51).

6.5.6 Polarity reversal and types of conditionals

We may now wonder whether non-counterfactual subjunctive conditionals would not be fine in the same contexts as indicatives are. A conditional like (46-a) can still be informative relative to contexts that satisfy the condition in (50), as long as the

context contains some $\neg\phi\text{-}\neg\psi$ -worlds. On the other hand, non-counterfactual subjunctive conditionals exhibiting polarity reversal would be uninformative relative to any context in which they would constitute a scalar assertion. Take, for example, the subjunctive conditionals in (44-b) and (45-b) and suppose there is a context that encodes the following regularity: no lights throughout the negative phase of John's not being in the office, lights throughout the positive phase of John's being in the office. In contrast to the case of indicative conditionals, in any context in which the subjunctive conditional exhibiting polarity reversal would constitute a scalar assertion, there can be no John-not-in-the-office and lights-on worlds (i.e., worlds where the antecedent holds but the consequent does not) given the assumption that the context encodes the regularity. Similarly, the selection function would deliver for any John-not-in-the-office world only worlds in which the consequent is true. So the conditional is not informative relative to such a context. But in the kinds of reasoning non-counterfactual subjunctive conditionals are employed for, the conditional premise has to be informative.

Thus there is an interesting asymmetry between indicative and non-counterfactual subjunctive conditionals, just as there is an asymmetry between counterfactuals and all other conditionals with respect to polarity reversal. Counterfactuals allow polarity reversal as a matter of course because the use of the conditional constitutes a scalar assertion in any context in which it is felicitous, and this is so because contextual entailments must be given up once the information that gives rise to them is revised. Indicative conditionals allow polarity reversal only relative to contexts with additional information that would guarantee that the conditional constitutes a scalar assertion. By contrast any context in which non-counterfactual subjunctive conditionals would constitute scalar assertions are contexts that defeat the purpose of their being used.

References

- Anderson, A. R. (1951) A note on subjunctive and counterfactual conditionals. *Analysis* 12(2): 35–8.
- Baker, C. L. (1970a) Double negatives. *Linguistic Inquiry* 1: 169–86.
- Baker, C. L. (1970b) Problems of polarity in counterfactuals. In J. Sadock and A. Vanek (eds), *Studies Presented to Robert B. Lees by his Students*. Edmonton: PIL Monograph Series 1, Linguistic Research Inc., pp. 1–15.
- Chierchia, G. (2006) Broaden your views: Implicatures of domain widening and the 'logicality' of language. *Linguistic Inquiry* 37(4): 535–90.
- Edgington, D. (1995) On conditionals. *Mind* 104(414): 235–329.
- Edgington, D. (2008) Counterfactuals. *Proceedings of the Aristotelian Society* 108(1, part 1): 1–21.
- Heim, I. (1992) Presupposition projection and the semantics of attitude verbs. *Journal of Semantics* 9(3): 183–221.
- Horn, L. (1989) *A Natural History of Negation*. Chicago: University of Chicago Press.
- Karttunen, L. (1971) Subjunctive conditionals and polarity reversals. *Papers in Linguistics* 4: 279–96.
- Karttunen, L. (1974) Presupposition and linguistic context. *Theoretical Linguistics* 1: 181–94.
- Kratzer, A. (1981) Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2): 201–16.

- Kratzer, A. (1989) An investigation of the lumps of thought. *Linguistics and Philosophy* 12(5): 607–53.
- Krifka, M. (1991) Some remarks on polarity items. In Dietmar Zaefferer (ed.), *Semantic Universals and Universal Semantics*. New York: Foris Publications, pp. 150–89.
- Krifka, M. (1995) The semantics and pragmatics of polarity items. *Linguistic Analysis* 25: 209–57.
- Ladusaw, W. A. (1980) *Polarity Sensitivity as Inherent Scope Relations*. New York: Garland Press.
- Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.
- Lewis, D. (1979) Counterfactual dependence and time's arrow. *Noûs* 13(4): 455–76.
- Lewis, D. (1981) Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic* 10(2): 217–34.
- Löbner, S. (1989) German *Schon—Erst—Noch*: An integrated analysis. *Linguistics and Philosophy* 12(2): 167–212.
- Schwarz, B. and Bhatt, R. (2006) Light negation and polarity. In R. Zanuttini, H. Campos, E. Herburger, and P. H. Portner (eds), *Crosslinguistic Research in Syntax and Semantics: Negation, Tense, and Clausal Architecture*. Washington, D.C: Georgetown University Press, pp. 175–97.
- Stalnaker, R. (1968) A theory of conditionals. *Studies in Logical Theory* 2: 98–112.
- Stalnaker, R. (1973) Presuppositions. *Journal of Philosophical Logic* 2(4): 447–57.
- Stalnaker, R. (1975) Indicative conditionals. *Philosophia* 5(3): 269–86.
- Szabolcsi, A. (2004) Positive Polarity—Negative Polarity. *Natural Language and Linguistic Theory* 22(2): 409–52.
- Tichý, P. (1976) A counterexample to the Stalnaker-Lewis analysis of counterfactuals. *Philosophical Studies* 29(4): 271–3.
- van der Sandt, R. (1991) Denial. *Papers from the Annual Meeting of the Chicago Linguistic Society: Parasession on Negation* 27(2): 331–44.
- Veltman, F. (2005) Making counterfactual assumptions. *Journal of Semantics* 22(2): 159–80.
- von Stechow, K. (1998) The presupposition of subjunctive conditionals. In U. Sauerland and O. Percus (eds), *The Interpretive Tract (MIT Working Papers in Linguistics 25)*, Cambridge, MA: MITWPL, pp. 29–44.
- Zeijlstra, H. (2012) Not a light negation. *Semantics and Linguistic Theory* 22: 545–64.

Counterfactuals and Probability

Robert Stalnaker

7.1 Introduction

Dorothy Edgington has long defended what Jonathan Bennett labeled the NTV thesis about conditionals: conditional sentences do not express propositions, but should be understood as expressions of conditional belief, where a conditional belief is qualified belief in the consequent proposition, with the antecedent proposition determining the qualification.*¹ Edgington developed her account of conditionals in a probabilistic context: conditional belief is conditional credence, or degree of belief, modeled by conditional probability. Her NTV thesis contrasts with the thesis, defended by David Lewis, among many others, that a conditional sentence expresses a proposition that is a function of the propositions expressed in the antecedent and consequent clauses (plus context).

Edgington has also long argued that the contrasting kinds of conditionals—standardly but inaccurately labeled subjunctive and indicative²—should receive a unified treatment. While all acknowledge that there are contrasts between indicative and corresponding subjunctive conditionals, including cases where the indicative is assertable, while the corresponding subjunctive is not (and vice versa), Edgington argues that in many cases, the subjunctive can be interpreted as a restatement, in a retrospective context, of what was said with an earlier indicative. More generally, she argues that we should explain the differences between the two kinds of conditionals within a common framework. On this issue, her account of conditionals contrasts

* Thanks to Lee Walters for very helpful comments on a draft of this chapter.

¹ For more on the NTV thesis see Bennett, 2003, Chapter 7.

² The ‘subjunctive’ label is grammatically inaccurate since while some so-called subjunctive conditionals in English do involve the grammatical subjunctive, it is tense and aspect rather than grammatical mood that is doing the work in marking the distinction. The label ‘indicative’ is also inaccurate, not only because some so-called subjunctive conditionals are literally indicative, but also because at least in archaic English, some conditionals that are literally (present tense) subjunctive should be grouped, by meaning, with the so-called indicative conditionals. (For example, ‘if he be found guilty, he will appeal.’) It is hard to find labels for the distinction that are both grammatically accurate and theoretically neutral (‘counterfactual’ is problematic for a different reason), but there does seem to be a robust distinction clearly marked in the grammar, and general agreement about paradigm cases of the contrast. I will follow tradition, as Edgington does, in using the subjunctive/indicative terminology, but these terms should be understood to mean ‘so-called subjunctive’ and ‘so-called indicative’. See Chapter 8 by Sabine Iatridou in this volume for an interesting discussion of grammatical issues involving conditionals, as well as an earlier paper, Iatridou 2000.

both with some propositionalists, such as David Lewis and Frank Jackson, who argue that indicative and subjunctive conditionals should receive fundamentally different propositional analyses, and also with theorists such as Allan Gibbard, who defend the NTV thesis for indicative conditionals, but a propositional analysis for subjunctive conditionals.³

On the first issue, I have been on the propositionalist side, while acknowledging that there is some, perhaps considerable, indeterminacy in the application of the truth-conditional semantics. That is, it is acknowledged that conditional sentences may often express only partial propositions. (If we identify a *proposition* with a function from possible worlds to truth-values, then a partial proposition is one that is defined for some but not all possible worlds.) On the second issue, I have been with Edgington on the unificationist side, aiming to explain the differences between the two kinds of conditionals as variations of a single semantic analysis. The two issues interact, since the NTV thesis is *prima facie* more plausible for indicative conditionals, while the propositionalist's case is strongest for subjunctive conditionals. So one may use the unification hypothesis, as Edgington does, to generalize the NTV thesis from indicative to subjunctive, or as I have, to motivate a propositionalist account of indicatives.

While I have been on the propositionalist side on the first issue, I have also argued that the distance between a conditional assertion account of indicative conditionals of the kind that Edgington has developed and the propositionalist account of indicative conditionals that I have defended is less than might appear. Specifically, I have argued that a version of the latter is essentially equivalent to a version of the former. One can see the conditional assertion account as a limiting case of a propositionalist account, and this helps to facilitate an explanation of the continuity between the two kinds of conditionals, and so to support a unified account. One of the things I want to do in this chapter is to develop further the ecumenical attempt at reconciliation, this time focusing on the contrasting accounts of subjunctive conditionals. I also will try to get clearer on exactly what is at stake in the debates about whether conditionals express propositions, and whether they have truth-values.

Here is my plan: In Section 7.2, I will review my attempt to reconcile a conditional assertion account of indicative conditional speech acts with a truth-conditional analysis of conditional sentences, connecting it with a more general project of reconciling a truth-conditional semantics for deontic and epistemic modals such as 'may', 'might', and 'must' with an expressivist account of the role of modal statements in discourse. But even if one *can* give the conditional assertion account of indicative conditionals a truth-conditional formulation, one may ask why one *should*—what advantages this kind of formulation offers. So in Section 7.3, I will try to motivate the truth-conditional formulation by appealing to continuities between the conditionals that best fit the NTV account and those that seem to be used to make factual claims. Edgington herself appeals to these continuities, using them to argue for extending the NTV account of indicative conditionals to counterfactuals, but I will argue in Section 7.4 that her way of connecting indicative conditionals with counterfactual

³ Gibbard 1980.

statements that are made in a later retrospective context is in tension with the thesis that indicative conditionals express conditional probabilities. Her defense of an NTV account of counterfactuals also appeals to the intuitive judgment that some counterfactuals should get probability values, but not truth-values. I agree with her intuitive judgments about the cases in question, and will try, in Section 7.5, to reconcile these judgments with a broadly truth-conditional analysis of counterfactual conditionals.

7.2 Expressivism about Indicative Conditionals and Epistemic Modals

The truth-conditional semantics that I proposed for conditionals in general posited a selection function takes two arguments (a possible world and a proposition) to a possible world.⁴ The intuitive idea was that if w is the actual world, then $f(w, \phi)$ is a world that differs from w only in the minimal way required to ensure that ϕ is true. Structural constraints were imposed on the function to ensure that it conforms to this idea, but no attempt was made, in the abstract semantics, to give any substantive criteria for determining what counts as a minimal difference, or a ‘closest’ possible world. That, it was assumed, depended on the application of the semantics, and on substantive metaphysical and epistemological hypotheses. My aim in proposing this semantic analysis was more modest than the aim of David Lewis, who independently proposed a formally similar semantic analysis at about the same time.⁵ Lewis intended his analysis (which was taken to apply only to the so-called subjunctive conditionals) to be part of a reductive project in the spirit of Hume, whom Lewis described as the great ‘denier of necessary connections’. I am skeptical of Lewis’s reductive project, and of the grand metaphysical thesis that he labeled ‘Humean supervenience’, but that is an issue that can be separated from the question whether a truth-conditional analysis is appropriate for conditionals.⁶

In contrast with Lewis, I took this kind of abstract truth-conditional analysis to apply to both indicative and subjunctive conditionals, but it was assumed that the grammatical marks that distinguish the two kinds of conditionals signaled a difference in the criteria by which minimally different possible worlds are selected. In the case of the indicatives, the criteria are epistemic, and I proposed the following constraint on the selection function: for possible worlds compatible with what is presupposed, the selected worlds must also be compatible with what is presupposed. That is, if C is the *context set*—the set of possible worlds compatible with what is presupposed in the relevant context, then a selection function f is *admissible* for the interpretation of an indicative conditional if and only if for any $w \in C$, $f(w, \phi) \in C$. (Since it is a general constraint on selection functions that $f(w, \phi) \in \phi$, the definition of an admissible selection function presupposes that indicative conditionals are interpreted only when the antecedent is compatible with the context set.) There will be many admissible selection functions. The constraint is just that the function

⁴ Stalnaker 1968 and Stalnaker 1984, ch. 7.

⁵ Lewis 1973.

⁶ I criticized Lewis’s reductive project in Stalnaker, 1984, ch. 8, and more recently in more detail in Stalnaker 2015.

or functions relevant to the interpretation of indicative conditionals must be admissible.⁷

I say function *or functions* because it was recognized from the start that in application, the context in which a conditional of any kind is interpreted may not fully determine the parameter relative to which the formal semantics specifies truth-conditions for the conditional. The idealized semantics makes a uniqueness assumption: for each proposition, there is a unique possible world (or possible situation) that is *the* possible situation in which the proposition is true, and that is minimally different from the actual situation (or more generally, the situation relative to which the conditional is being evaluated). But in practice, the relevant context may provide only constraints on the parameter that do not fully determine it. The gap between ideal semantic theory and practice is not a special feature of conditionals: any contextual parameter (a domain of discourse for the interpretation of a quantifier, a reference class for the interpretation of a gradable adjective) might be only partially determined. A supervaluation strategy⁸ is used to model this kind of gap between theory and practice: a statement is true if true for all precisifications of the partially determined parameter, false if false for all, and neither true nor false if true for some and false for others. The supervaluation strategy allows compositional semantics to proceed at the level of ideal theory, with supervaluations determined for the result.⁹

If we have models that allow for a class of selection functions, rather than a single selection function, then we can consider the following limiting case of my account of indicative conditionals: let the class of selection functions be the class of *all* selection functions that are admissible, relative to the given context set. In a model of this kind, a conditional will be true in a possible world compatible with the context if and only if the consequent is true in all possible worlds in that context in which the antecedent is true, false if false in all of those possible worlds, and neither true nor false otherwise.

⁷ Stalnaker (1975).

⁸ See van Fraassen (1966), where the supervaluation idea is first introduced.

⁹ Supervaluations are just one part of a multi-pronged strategy for reconciling the uniqueness assumption in the formal semantics with the manifest fact that it is hopelessly implausible to assume that we have what the ideal theory seems to require: a well-ordering of all possible worlds relative to each possible world. The implausibility is lessened when one recognizes that the semantics need not assume that the entities labeled 'possible worlds' are metaphysically complete ways a world might be, but can be relevant alternative possibilities, individuated by what is at issue in a particular application. One might think of the possibilities as partition cells of a partition of logical space, rather than point in the space, and where the partition is relatively coarse-grained, it not so implausible to suppose that we have a function that selects a unique one that is minimally different from the actual situation. But it should also be said that we need not assume that even in such a limited context, conditionals are defined for all possible antecedent suppositions. One may think of the antecedent of a conditional as something like a definite description. The speech act of supposing that ϕ might be thought of as something like a request to consider *the* minimally different possible situation in which ϕ . The supposition is appropriate (just as a singular definite description is appropriate) only in contexts that provide a unique salient situation, or referent.

The need to jump through various different hoops to save the semantics that validates the principle of excluded middle raises a methodological concern: are these maneuvers just ad hoc devices to save a theory? But whether one should worry about that depends partly on whether the linguistic evidence supports the validity of the principle of excluded middle, as I think it does, and also on whether the necessary maneuvers have independent interest and motivation, which I think they do.

This limiting case is the maximally cautious, or minimally committal interpretation that fits the truth-conditional analysis of indicative conditionals sketched above. If we combine this truth-conditional semantics with an account of speech act force, we get a general account that is essentially equivalent to a conditional assertion account. To make an assertion, on the general account of assertion that I have given, is to propose that the propositional content of the sentence asserted (in the relevant context) be added to the common ground. With ordinary assertions that purport to convey factual information, the sentence is interpreted in the prior context, and the proposition determined relative to that context then is added to determine a posterior context (assuming the assertion is accepted). But in some cases, including deontic and epistemic modals, one may interpret the sentence in a prospective way, relative to the *posterior* context. The speech act rule will be something like this: adjust the context in the minimal way required to make the sentence true relative the resulting context. The truth-conditional semantics plays an essential role in the account of speech acts of this kind, but the result helps to explain their expressive or practical role: the sense in which a statement does not make a factual claim, but expresses an attitude other than categorical belief. So, for example, one can take a statement like ‘you may have a piece of pie’, not as stating the prior fact that having a piece of pie is permissible, but as a speech act of giving permission. One can understand ‘the butler might have done it’ as an expression of the fact that the speaker is not in a position to rule out the possibility that the butler did it. The ‘might’ statement, on this account, is an expression of the fact that the speaker is not in a position to rule out the possibility, but it is not a statement of that fact. It can be understood as a proposal that the possibility not be excluded from the context. If we take indicative conditional assertions to be speech acts of this kind, with the truth-conditional semantics I have sketched, then we have an account that can be seen as a formulation of a conditional assertion analysis: an account that says that one is in a position to assert (if ϕ , ψ) if and only if one is in a position to assert ψ in a derived context obtained by adding the condition ϕ to the prior context. In making a conditional assertion, one is expressing a conditional belief.¹⁰

Of course, beliefs and conditional beliefs can be partial, as Edgington’s work emphasizes. Her account began with Adams’s thesis: that a conditional is assertable if and only if the conditional credence of consequent on antecedent is sufficiently high. I will look at some putative counterexamples to Adams’s thesis in Section 7.3, but for now I want just to make two points about the relation between Adams’s thesis and our truth-conditional formulation of a conditional assertion account. First, our account of conditional assertion does not *entail* Adams’s thesis: the conditional assertion account says that one is in a position to assert ‘if ϕ , then ψ ’ in a context C if and only if one is in a position to assert ψ in a derived context defined by adding (temporarily) the supposition ϕ to the context C , but nothing is said about the conditions in which one is in a position to make an assertion in a given context, either a basic context or a derived one that is created by a supposition. The hypothesis

¹⁰ See Stalnaker (2014), ch. 6 for discussion of this kind of account of deontic and epistemic modals, and indicative conditionals. The basic idea starts with David Lewis’s game of commands and permissions. See Lewis (1975).

that conditional assertion goes by conditional probability is an additional element of the theory, one that adds a Bayesian norm of assertion.

The second point is that the truth-conditional formulation of the conditional assertion account of indicative conditional statements is *compatible* with Adams's thesis. That is, the truth-conditional analysis is compatible with a further hypothesis that one is in a position to make a conditional assertion if and only if one's conditional credence of consequent on antecedent is sufficiently high. One might be tempted to think that David Lewis's notorious triviality proofs¹¹ showed that one cannot consistently combine Adams's thesis with a propositional analysis of the conditional, but this is not correct. Lewis's proof (and most of the many variations of it) was a *reductio* with two assumptions: (1) that conditionals express propositions, and (2) that Adams's thesis holds for a nontrivial class of probability functions that is closed under conditionalization. (A class of probability functions is defined as nontrivial if it includes at least one probability function that assigns non-zero probability to at least three disjoint propositions.) But there was also an implicit assumption, which Bas van Fraassen identified, and labeled 'metaphysical realism'.¹² The implicit assumption is an invariance condition: roughly, that the proposition expressed by a conditional sentence does not vary with a change in the context, where a credence function is taken to be an essential part of the context. Conditionalization is a shift from one credence function to another, and so from one context to another. The shift is from a prior belief state to the posterior state one would move to upon receiving the information that is the content of the condition. Suppose we label the prior credence function 'P', and the posterior credence function that would result from learning B and conditionalizing 'P_B'. Assume that the arguments of the probability functions are *propositions*—the ones expressed by the sentences, A, B, $A \rightarrow B$, etc. Then for any proposition X, $P_B(X) =_{df} P(X/B)$. A crucial step in Lewis's argument assumes that $P_B(A \rightarrow C) = P(A \rightarrow C/B)$. This looks like a straightforward application of the definition, but it assumes that the conditional, $(A \rightarrow C)$ expresses the same proposition in the prior context as it expresses in the posterior context. If this assumption is not satisfied, then the argument will suffer from a fallacy of equivocation. And it is clear, given the pragmatic constraint that our truth-conditional analysis imposed on the interpretation of indicative conditionals, that the invariance assumption will not hold. If what is accepted in a context changes, then different constraints are imposed on the interpretation of the indicative conditional, and this will require that in some cases, that a different proposition is expressed in the posterior context.

Van Fraassen not only identified a questionable implicit premise of Lewis's proof of his triviality thesis, but also proved a counter-thesis—a possibility result. What Van Fraassen proved was this: for any probability model defined on a set of sentences of an extensional propositional language, one can extend it to a probability model for the richer language that adds a conditional connective, and in which Adams's thesis is satisfied. That is, for any two sentences A and B of the original language for which $P(A) \neq 0$, $P(A \rightarrow B) = P(B/A)$.

¹¹ Lewis (1976). See Hájek and Hall (1994) for an excellent discussion of Lewis's result, and related issues.

¹² van Fraassen (1976). See also Stalnaker and Jeffrey (1994).

7.3 From More Fragile to More Robust Conditionals

So one *can* give a truth-conditional form to an expressivist account of indicative conditionals—one that takes conditional speech acts as expressions of conditional belief, as represented by conditional credence. More generally, one can reconcile truth-conditional analyses of deontic and epistemic modal expressions with an account of their role in discourse that interprets them as doing something other than making factual claims. But one may ask—now focusing just on the case of conditionals—what advantage does this kind of formulation have? The response to the triviality arguments bring out the fragility of the propositions that are expressed, according to this formulation, by conditionals that conform to Adams's thesis. The model that van Fraassen constructed for his possibility proof is just one way of reconciling Adams's thesis with a truth-conditional analysis, but any way of constructing a model that does this job will have to define the model in term of a given credence function, and when the credence function is changed, one will have to redefine the model if Adams's thesis is to hold also relative to the new credence function. And it is not just that the interpretation of the conditional must be highly context-dependent (that what proposition is expressed depends on context) but that the propositions themselves seem not to be detachable from the epistemic context in which they are expressed. That is, it is not clear how we can express, in a new context, after one's epistemic situation changes, what was said before with a conditional assertion.

The upshot is that the triviality results, and the response to them, provide a theoretical argument for the conclusion that any propositional analysis of conditionals that is compatible with Adams's thesis will necessarily be highly fragile: the proposition expressed by a conditional will shift with any shift in the epistemic state of the parties to the context. There are also intuitive examples that help to make this point, the most famous of which is Allan Gibbard's story about a poker game on a Mississippi riverboat:¹³ Mr. Stone has raised, and it is up to Sly Pete to call or fold. Jack has seen both hands and knows that Pete has a losing hand. Zack has seen only Mr. Stone's hand, which is quite good, but he has signaled its content to Pete, so he knows that Pete will call only if he has a hand that can beat Mr. Stone's. At this point the room is cleared, and Jack and Zack convey their information to a third person, Allan. Jack says, in a note to Allan, 'if Pete called, he lost', and Jack says, in a separate note, sent independently, 'if Pete called, he won'. Allan accepts both statements, and concludes that Peter folded.

The example has the following structure: There are three relevant alternative possibilities, X, Y, and Z. One person has information that excludes possibility Z, and a second person has information that excludes possibility Y. The two informants convey their information to a third person independently by making conditional assertions. What our truth-conditional formulation of the conditional assertion account says about the example is that the conditional statements made by the two informants must be interpreted relative to different contexts. Gibbard carefully sets up the example so that the *prior* context of each conditional assertion is the same: it is a context in which all three possibilities are compatible with what is presupposed.

¹³ Gibbard (1980).

But our truth-conditional account treats the conditional assertion as a proposal to change the context to ensure that the sentence, as interpreted in the *posterior* context, is true. The context, as changed by Jack's statement, is different from the context, as changed by Zack's statement. Since the actual situation—the situation in which Pete knew he had a losing hand, and so folded—is compatible with what both Jack and Zack said, our truth-conditional account will say that both of their statements were true, which implied that the proposition Jack affirmed must be different from the one that Zack denied.

One might think that the point of giving a propositional analysis is to identify an item of information which can be detached from the context in which it is expressed, and from the epistemic situation of a person who believes it, an item that might be assessed, in retrospect, in a later context. But this does not seem possible with examples that fit the pattern of the Sly Pete story. Imagine Allan thinking to himself, after getting the messages from his informants (but still not having gotten definitive information about how the game ended), 'suppose Pete did call after all. Did he then win or lose? If he did call, then one of my informants was mistaken. Perhaps *Jack* was right, but Pete missed Zack's signal, and so mistakenly thought he had a winning hand. Or alternatively, perhaps Zack was right, but Jack, in observing the two hands, mistook a jack for a king, or a spade for a club, and so was wrong that Pete had a losing hand.' It seems intuitively clear that Allan might think this way, even if he had complete confidence that both of his informants were correct in what they said. Suppose he takes himself to know, based on what his informants told him, that Peter folded. He can still consider what he would think if something he took himself to know turned out to be false. If this is the situation, then the question he is asking himself cannot be a question about the truth or falsity of propositions expressed by Jack and Zack. He is asking not 'which of my informants was wrong?' but rather, 'if one of them is wrong (because, contrary to what I take myself to know, Pete did indeed call), which one of them is it?' So Allan's conditional question is not a retrospective assessment of Jack's and Zack's statements. It is also clear that Allan's question is not the counterfactual question, 'if Pete had called, would he have won or lost?' Assuming as we do that Allan trusts his informants, the answer to this question is clear: he would have lost. This does not show that Zack was wrong when he said, 'if Pete called, he won' for the following reason: Zack was well aware that he was not in a position to rule out the possibility that Pete had a losing hand (in which case he would have lost if he had called), and yet he was in a position to say that if Pete called, he won. So it can't be that his statement ruled out this possibility.

If conditionals of the kind that are involved in the Sly Pete story were the only ones, there would be little point in giving a truth-conditional formulation to the conditional assertion account, but with some indicative conditionals, and certainly with counterfactual conditionals, it does seem that we can identify something like an item of information that is not tied so tightly to the context in which it is expressed. In the Sly Pete story, there is no disagreement between Jack and Zack, even though their statements seem, on the surface, to make incompatible claims. It is just that each has only partial information, which when they share it with each other will change their epistemic situations so that the question of retrospective assessment of their apparently contrary claims does not arise. But the situation with other examples,

including some that share the abstract structure of the Sly Pete story, will be different. Here is an old example I have discussed before, based on a case that Paul Grice described in his William James lectures in 1967:¹⁴ There are three candidates, Wilson, Heath, and Thorpe. Alice is convinced that if the victor is not Wilson, it will be Heath, while Bert thinks she is wrong about this—if it is not Wilson who wins it will be Thorpe. The case is like the Sly Pete case in that there are three possibilities and two agents, one who takes herself to be in a position to rule out one of the three possibilities, while the other takes himself to be in a position to rule out a different possibility. But in contrast with the Sly Pete story, in this case the two hold their ground, each rejecting the other's claim. They take themselves to be disagreeing, and each may defend his or her claim against the other, presenting evidence that he or she is right. As they debate the issue, the context keeps shifting, both introducing new information that they take to support their case. The disagreement persists, and even though the context keeps changing as new information about polls, precedents, or policy positions is introduced, Alice and Bert take themselves to be addressing the same question throughout, and to be defending contrary answers to it. Even after the election, which Wilson wins, the disagreement may persist, with the conditional now taking a counterfactual form. Alice insists she was right—the election might have gone the other way, and if Wilson hadn't won, it would have been Heath. Bert agrees that the result might have been different, but remains convinced that if the winner hadn't been Wilson, it would have been Thorpe.

The fact of disagreement does not by itself require the conclusion that the conditional aims to make a factual claim that is true or false. Two people may disagree about what they are in a position to accept, or what beliefs it is reasonable for them to have, given their shared evidence, even if there is no factual claim that one regards as true and the other regards as false. But the dialectic of disagreement, and the retrospective assessment, in a later context (after more information has come in) of what was said before, suggest that there is something like an item of information in dispute that persists through a shifting context, an item that a truth-condition helps to identify. Conditionals may begin simply as expressions of hedged commitments, reflecting holistic properties of one's current epistemic situation. Some conditional beliefs are ephemeral, reflecting parochial features of one's particular epistemic situation—what one happens to know and be ignorant of. The Sly Pete story was designed to be a case of this kind, a case where there is no retrospective question: which of the two apparently conflicting conditional claims—Jack's or Zack's, was correct. But in general, we have reason to try to find more stable or robust conditional beliefs that can be generalized and applied in a range of different particular epistemic situations. Sometimes it is appropriate to project one's epistemic priorities onto the world, to hypothesize that the world is such that one would be (under normal conditions) in a position to infer ψ upon learning ϕ . But in a particular case, it may be uncertain or controversial whether a dispute about a conditional is one that would be settled by a fact that is independent of the truth or falsity of the supposition or whether the disagreement just reflects a difference in

¹⁴ See Grice (1989: 64), and Stalnaker (1984: 113).

the epistemic perspectives of the parties to the dispute. In the British election example, retrospective assessment after more of the relevant facts are in might reveal that the dispute about Wilson, Thorpe, and Heath was really more like the Sly Pete case in that the right explanation for the disagreement was that Alice had information that, properly understood, was sufficient to rule out Thorpe, while Bert had information that sufficed to rule out Heath. But on the other hand, a retrospective assessment might reveal that there are factual questions on which their dispute turned that settle the question, if Wilson hadn't won, which of the other two would it have been?¹⁵ Given that there may be, at the time of a dispute, uncertainty about whether there is a fact of the matter, and given that there is continuity between the more ephemeral conditionals that are essentially tied to a specific epistemic situation and more robust conditionals that can be detached from their epistemic contexts and assessed against the facts, it is useful to have a semantic account of conditionals that covers the whole range of cases. A truth-conditional semantics that allows for both context-dependence and truth-value gaps is one that meets this condition.

7.4 Retrospective Assessment

The points made in the last section about the continuity between indicative and counterfactual conditionals, and about retrospective assessment of indicative conditional claims, are not matters of dispute between my account and Edgington's. In fact, she argues forcefully and persuasively for exactly this kind of continuity, using it as part of her argument for extending the non-truth-conditional analysis of indicative conditionals to the case of counterfactuals. She makes a convincing case that a counterfactual conditional may, in some cases, be understood as a restatement of what was said earlier, in a different context, with an indicative conditional. But there is a tension between this observation and Adams's thesis that conditional probabilities provide the assertability conditions for indicative conditionals. Let me first review one of her striking examples, and then say what I think the tension is.

According to Edgington's story,¹⁶ our protagonist (call her Dorothy) is driving to the airport when her car breaks down, and as a result she misses her plane to Paris. This was fortunate, as it turned out, since the plane crashed and everyone on board was killed. Coincidentally, a fortune teller had predicted this disaster, saying to Dorothy, 'If you take that plane, you will be killed.' Dorothy put no stock in the fortune teller's prediction. She ignored it, and when she learned of the crash she rightly took it to be a ghastly coincidence. But still, when she learns of the crash, she says of the fortune teller, 'My God, she was right, if I had taken that plane I would have been killed.'

¹⁵ The question is not just whether there is a counterfactual question, but whether the answer to the counterfactual question counts as a retrospective assessment of the original epistemic dispute. As I argued above, in the Sly Pete story there is a clear correct answer to the counterfactual question, 'if Pete had called, would he have won?' is not relevant to judging whether Jack or Zack was correct in their epistemic conditional judgments.

¹⁶ Edgington (2004: 12).

Edgington uses this story to make a number of points, but the one I want to focus on here is that what is said with the counterfactual is taken to be the same thing as what was said earlier by the fortune teller with an indicative conditional. Edgington does not think that the claim that the fortune teller was right is the claim that what she said was true, since on her account conditionals don't have truth-values, but she is observing that there is some kind of item that was expressed in the earlier context, and that can be expressed again, and assessed later, in a different context. This seems right to me, but it raises a problem for the hypothesis that the indicative conditional expresses a conditional belief, as represented by a conditional probability. The problem is that the appropriateness of the retrospective assessment of the indicative conditional suggests the following constraint on indicative conditionals—a kind of reflection principle: one should not make a conditional assertion if one has good reason to believe that a retrospective assessment would judge the assertion to have been mistaken. This constraint is not a problem for the fortune teller example, where even though Dorothy later judges correctly that the indicative statement was right, neither she nor the fortune teller had reason to think this in advance. Since there was at the time reason to think that the conditional claim would be seen in hindsight to be mistaken, the reflection principle implies, as one should expect, that the statement was unjustified. But there are other examples that show a conflict between what the conditional credence prescribes and what is required by the anticipation of a retrospective judgment. These include examples that have been posed in the literature as counterexamples to Adams's thesis. The counterexamples are controversial, but I want to suggest that Edgington's point about retrospective assessment provides some theoretical rationale for the judgment that the examples are indeed counterexamples.

Start with a story told by Vann McGee:¹⁷ a contestant on a television game show ('To Tell the Truth') is believed to be Sherlock Holmes. It is not certain that he is Holmes—say that our observer is about 90% sure that he is. A man, Murdock, has died under suspicious circumstances, and before hearing what the game show contestant says, the observer was inclined to believe that the death was probably an accident. But the putative Holmes asserts that it was murder, and that he is almost certain that it was Brown who did it. He further asserts that it was definitely murder, so if it wasn't Brown who did it, it was someone else.

Since our subject is 90% sure that the contestant is indeed Holmes, and since he takes the perceptive Holmes to be authoritative on questions about the causes of suspicious deaths, he is highly confident that what the contestant said is true: it was murder, Brown did it, but if he didn't someone else did. Should he learn, however, that Brown did not do it, this would cast doubt on the judgment that the contestant is really Holmes, since it is unlikely that Holmes would be both confident and mistaken about who did it. But if the contestant is not the real Holmes, then there is no reason to believe that this is a case of murder at all. So our subject's conditional credence—that someone else did it, on the condition that Brown did not—is much lower than his belief in the conditional asserted by the contestant—that if Brown didn't do it, someone else did.

¹⁷ McGee (2000).

If we apply the retrospective reflection principle to Holmes's conditional claim, we get some support for McGee's intuitive judgments about this example. Our subject is 90% certain that the contestant is Holmes. What he therefore expects (with 90% certainty) is that when the truth about the contestants is revealed at the end of the show, we will learn for sure that the contestant is indeed Holmes. If that happens, he will then have strong reason to believe that Murdoch's suspicious death was murder, that Brown did it, and that if he didn't, someone else did. At that point (if what our subject expects with 90% certainty does transpire), credence in the indicative conditional will coincide with the *revised* conditional credence, but at the earlier point, it is the *anticipated* conditional credence (what it is expected to be when the facts about the contestants comes in) rather than the conditional credence at that time that determines the credence in the conditional. Just to spell it out: let H be the proposition that the contestant is Holmes, M the proposition that Murdoch was the murderer, and E the proposition that someone else was the murderer. Let P be the subject's initial credence function and let P_H be the anticipated credence function—the one that the subject expects to have after learning H . Then while $P(E|\sim M)$ is not high, $P_H(E|\sim M)$ is high. Since the subject expects that he will have reason, in retrospect, to accept the conditional, he has reason to accept it in the initial state.

It should be noted that this example is not a case, like Edgington's plane crash case, where the retrospective judgment about the conditional is made with a counterfactual conditional. As McGee pointed out, in this case, there is a clear contrast between the indicative conditional statement (both at the earlier point, and at the point of retrospective assessment, after then identity of the contestant is revealed) and the corresponding counterfactual. Neither Holmes nor our subject is inclined to give any credence at all to the claim that if Brown hadn't done it, someone else would have. It might be a robust part of the theory about the crime that whoever did it acted alone, so that if the actual murderer (whoever it turns out to be) had not done it, the deed would not have been done. But there are other putative counterexamples to Adams's thesis where the anticipated retrospective judgment is made with a counterfactual conditional. The examples posed by Stefan Kaufmann¹⁸ fit this pattern. Since these examples are more abstract than McGee's, involving balls drawn from urns, they also reveal more clearly the structure of the cases, and provide hard numbers for the probability judgments.

Suppose a ball is to be drawn at random from one of two urns, A or B . We are 90% certain that it will be from urn A , whose composition is as follows: 49 white balls, and one red ball with a black spot. Urn B also contains 50 balls, one white, and 49 red, all without a black spot. What is the right credence, given these facts about the case, for the conditional, if a red ball is drawn, it will have a black spot? There is some inclination to say that it is .9, since we are 90% sure that the urn is A , and a red ball will have a black spot if and only if it is drawn from urn A . But the conditional probability of a black spot, given a red ball, is only about .155. For if a red ball is drawn, it is very probably from urn B , in which case it will lack the black spot.

What should one expect to happen in this case? It is almost 90% certain that a white ball will be drawn from urn A , in which case it will then be certain that if a red ball had been drawn (unlikely as that would have been), it would have had a black

¹⁸ Kaufmann (2004).

spot. If the counterfactual judgment is reasonably taken to be a retrospective judgment of the indicative conditional, then the person who claimed earlier that probably, if a red ball is drawn it will have a black spot will be vindicated. So since in the initial context, one gives high credence to this outcome, the reflection principle implies that one should at that point affirm this conditional (or judge it to be highly probable), and not the contrary one that goes by the conditional probability.

One's intuitions about this kind of case may be pulled in two directions. Some are inclined to think that one should still assess the indicative conditional by the conditional probability, and that intuitions to the contrary are based on a cognitive illusion. But there is more at stake here than just intuitions about examples. The putative counterexamples are all cases where the conditional credence is unstable in the sense that it would be easily changed, in one direction or the other, by salient new information, information that the agent knows would result in a more stable conditional probability judgment. The probability space is naturally partitioned into alternative hypotheses (the contestant is Holmes, or not; the urn is A, or B) that meet the following condition: Learning which of these alternative hypotheses is true would stabilize the conditional credence, and give it an objective rationale. One thus has reason to be interested in conditionals that express, not just the current disposition to change one's credences, but a more stable disposition to change one's credences that one *would* have if one learned certain salient facts. Such conditionals may help to give more structure to one's epistemic situation, and to identify epistemically valuable questions to ask in one's inquiries. The moves from indicative to subjunctive conditionals, and from subjective credence to objective chance are moves that aim to find robust reflections of epistemic priorities in the judgments of fact about the world.

The truth-conditional semantics provides a formal structure that helps to model this kind of move from conditionals that express local and ephemeral features of an epistemic situation to conditionals that aim to reflect robust features of the world. The pragmatic constraint on indicative conditionals gives a class of admissible selection functions, and in the limiting case—the most cautious and noncommittal kind of model—all admissible selection functions are on a par. In this kind of model, conditionals are true only if true relative to all admissible selection functions, and false only if false for all. Indicative conditionals, on this limiting interpretation, do no more than express the epistemic disposition to accept the consequent upon coming to accept the antecedent. The speech act is a proposal that this epistemic policy be adopted for the moment. But the admissibility constraint is also compatible with structure that further constrains the selection functions. The theory allows for a partition of the space of possibilities, with the selection function constrained to select a possible situation from within its partition cell. That is, if the partition is $X_1, \dots, X_n$, then it might be required that for any proposition ϕ that overlaps X_i , and for any $w \in X_i$, $f(w, \phi) \in X_i$. This gives plausible models for McGee's 'To tell the truth' example, and for the urn examples discussed by Kaufmann. The more fine-grained the partition, the more the conditional is making a factual claim on the world.¹⁹

¹⁹ Brian Skyrms developed, some time ago, formal machinery for modeling this kind of process. See Skyrms (1980, 1984).

The indicative conditional, because of the pragmatic admissibility constraint, remains epistemic, but where we find a more stable partition, it sometimes allows us to detach the factual claim implicit in the conditional judgment from uncertainty about the truth of the antecedent. The tense/aspect/mood morphology that we inaccurately label 'subjunctive' signals that some of the presuppositions that define the speech context are being temporarily suspended, allowing us to say something like this: Independently of our actual knowledge or ignorance of the truth of the antecedent, the facts are such that (under normal conditions) if we didn't know whether the antecedent was true or not, we would be in a position to accept the consequent upon learning that the antecedent was true.

7.5 Chance and Counterfactuals

There seems to be no disagreement between Edgington's account of conditionals and the one I am promoting on the following two points: First, there is an item of some kind—something like a piece of information—that is expressed by a conditional, and that in some cases can be expressed and assessed in different contexts, including retrospective contexts in which what was earlier said with an indicative conditional is said with one that is counterfactual. Second, there is a distinction between judging such an item to be reasonable or justified and judging it to be *correct*. Edgington's plane crash example, with its prediction by the fortune teller, was designed to provide a dramatic illustration of this contrast—a case where a speaker was completely unreasonable and unjustified in making a conditional assertion which nevertheless turned out to be correct. One might think that the essential point of the thesis that conditionals express propositions is the claim that they express items of information that can be detached from the context in which they are expressed, and assessed as correct or incorrect independently of whether believing them was reasonable or unreasonable. But Edgington aims to reconcile these points about conditional statements and beliefs with a non-truth-conditional account. She writes, 'we are all familiar with the thought that rationally held beliefs may turn out false, and, conversely, something which there is no reason to believe may turn out true.... If that were my story [about the "correct"/reasonable distinction in the case of the conditional], there would be no novelty or mystery. But that is not my story. Counterfactuals, like other conditionals, are believed to the extent that a certain conditional probability is judged to be high, and that is not the probability of the truth of a proposition.'²⁰ In supporting her resistance to a truth-conditional analysis, she gives a variation of the plane crash example in which the fortune teller's prediction was hedged: she said 'I'm pretty sure that if you fly this week you will be killed', and then it turned out to be 90% of the passengers, rather than all, who are killed in the crash of the plane. Dorothy's counterfactual judgment, made in retrospect, was that 'it was very likely that I would have been killed, had I caught that plane'. This retrospective probabilistic judgment is 'right' even if there is no fact of the matter whether she

²⁰ Edgington (2004: 22).

would have been one of the 90% who were killed, or instead one of the 10% who survived. The conditional probability, she maintains, is not a probability of truth.

David Lewis's truth-conditional semantics for counterfactuals would judge that the counterfactual, 'If I had caught the plane, I would have been killed' is determinately false in the case where there is only a 90% (or 99.9%) chance that the speaker would not have been among the few survivors. Edgington resists this conclusion; as she says, Lewis's truth conditions 'make it too easy for a counterfactual to be plain false'.²¹ I agree with her about this, and I agree with her conclusion that we should judge the counterfactual to be highly probable, even though we may also say that there is no fact of the matter about whether it is true. As discussed above, the truth-conditional semantics I have defended allows for indeterminacy in application, with both indicative and subjunctive/counterfactual conditionals. So our truth-conditional approach can allow that there is no fact of the matter about whether a counterfactual such as the one in the modified plane crash case is true or false. But we will have a satisfactory account of this kind of case, and a plausible reconciliation of a truth-conditional theory with Edgington's approach, only if we can agree that counterfactuals without truth-values can have well-defined probability values. In most familiar cases of semantic indeterminacy, and of a supervaluation strategy for modeling them, we don't get probability values for the indeterminate cases. One does not normally judge, of a borderline case of baldness, that the person is 70% likely to be bald. But there is a different kind of indeterminacy, at least according to some metaphysical views, where probability judgments of propositions that are thought to lack truth-value seem to be natural. Consider the metaphysical thesis that there is no fact of the matter about future contingents. On this picture, the past and present are settled, but the future is open, and all possible futures are on a par in the sense that there is (now) no fact of the matter about which of the alternative futures will be realized. Still, this metaphysical view allows one to make predictions, and to speculate about how things will turn out. A semantics for a language appropriate to this metaphysical theory will be based on a branching tree structure where the nodes of the tree represent possible present moments, and total paths through the tree represent possible histories. Though there is, at any time, no fact of the matter about which total path through the tree is the actual one, the semantics can assign truth-values to the sentences of the language relative to the total paths (possible histories), and one can do the compositional semantics on propositions that are defined as functions from histories to truth-values. So the statement, 'there will be a sea battle tomorrow' might be true relative to some total histories that share our present moment and false relative to others, and therefore (applying the supervaluation method) neither true nor false *simpliciter*. The temporalist metaphysics allows that one might make a speculative prediction—'there will (I conjecture) be a sea battle tomorrow'—even when one knows that there is now no fact of the matter about whether one's prediction is true. Predictions are made to be assessed when the time comes.²²

²¹ Edgington (2004: 14).

²² See Thomason (1970) for a semantics of this kind, and MacFarlane (2003) for a discussion of this kind of temporalist metaphysics.

Now the key point for our purposes is that even if the proposition that there will be a sea battle tomorrow is now neither true nor false, that proposition might now have a probability value. Suppose, for example, that it is a fact about the present that there is a 75% chance of a sea battle. For this kind of more metaphysical indeterminacy, it is natural to take propositions that lack truth-values as having probability values. Is this probability of truth? It is probability of truth relative to a history.

You don't have to sign on to this metaphysical theory (as I do not) in order to find it intelligible (as I do) and to use it as a kind of precedent for a case where the thesis of metaphysical indeterminacy may be less controversial. Suppose we have a chance model, based on exactly the kind of tree structure used to model the story told just above. Attached to each branch from each choice point in the tree is a probability value—the chance, at that point, that that branch will be taken. The model is neutral about the metaphysical question just considered. One might be a metaphysical realist about the future, holding that there is a fact of the matter about which total history is actual, even though the fact is not determined by the present state of the world: on this realist account, future contingent propositions are true (now) in virtue of the way things will in the end turn out. Alternatively, one might sign on to the above temporalist metaphysical picture. Either way, each choice point in the model determines a class of histories, and one can define propositions as functions from histories to truth-values. On either metaphysical view, there will be a fact of the matter about what the objective probability of a proposition is at any given point. On the realist interpretation, a proposition about a future contingent might be both true, at a certain moment (because it is true relative to the history that will turn out to be actual), and also have (at that point) only a 10% chance of being true.

So the semantics is metaphysically neutral. The temporalist and the realist both have reason to define propositions in terms of total histories, even though the temporalist thinks that there is no fact of the matter about which history is actual. In order to get the compositional semantics to work smoothly the temporalist takes the points of evaluation as something more fine-grained than what her metaphysics claims is determined by the facts, but making this move is metaphysically innocent, since she need not use any resources in her semantics that her metaphysics does not provide.

What I want to suggest is that we can make a similar move, this time to points of evaluation that are even more fine-grained than total histories, in order to get a smooth compositional semantics for counterfactuals. Define a *choice function* as a function that determines not only a unique path through the tree, but also a choice at each choice point in the tree, including those on paths not taken.²³ Suppose we take our 'propositions' to be functions, from total choice functions, rather than just total histories, to truth-values.²⁴ This move, like the temporalist's, is metaphysically innocent, since all the material needed to define choice functions is available even if we deny, as seems plausible, that there is a fact of the matter about which choice function is the actual one. What this move buys us is this: first, there is a natural

²³ In addition, a choice function may order the choices at choice points with more than two alternatives.

²⁴ See Thomason and Gupta (1980) for a semantic theory for conditionals and branching time that develops this strategy.

comparative similarity relation that orders the choice functions, relative to each choice function that can be used to define a selection function that will yield a unique nearest 'possible world' (choice function) relative to propositions that are plausible antecedents for counterfactuals. Second, there is a natural probability distribution on the space of choice functions that will determine probability values for the counterfactuals. Counterfactuals that are true relative to some choice functions determined by a given history and false for others will have no truth-values, but well-defined probability values.

There is one further twist to be added to the story before we have completed what I hope is a reconciliation of Edgington's account with mine. We need to ensure that our chance structures can model a crucial distinction that Edgington emphasizes in her discussion of conditionals in hindsight—a distinction that involves the notion of causal independence. The temporal order implicit in the chance model already captures one constraint on the notion of causal or counterfactual independence: it is presupposed that earlier events are independent of later ones in the sense that if a later chance event had been different than it in fact was, the earlier one would still have been the same as it was. But this is a sufficient and non-necessary condition for causal or counterfactual independence. Edgington contrasts these two kinds of cases: (1) I pick a coin from a bowl of fair coins and flip it: it lands heads. If I had chosen a different coin instead, there is no reason to think it would have landed heads: the chance that that would have happened is 50/50. (2) Two coins are selected from the bowl and separately flipped, perhaps in different rooms, one slightly later than the other. Both landed heads. This time it seems reasonable to say that if the coin that was flipped first had landed tails instead, the other coin would still have landed heads, as it actually did. The second kind of case has been labeled a *Morgenbesser* case, after an example attributed to Sydney Morgenbesser: I decline to bet on the flip of a coin, which is then flipped, landing heads. I say, 'If I had bet on heads, I would have won', and this seems obviously right. It is also true that if I had bet on heads, at that point in time, there would have been only a 50/50 chance that the coin would land heads. But if, as is plausible, the flipping event was causally independent of the betting decision, the counterfactual statement will be correct. *Morgenbesser* cases, where later events are causally independent of some earlier ones, play a crucial role in Edgington's account of the hindsight judgments, where earlier indicative conditional claims are vindicated by later counterfactual judgments.

We can account for this distinction in the context of our tree-structure models by recognizing that distinct nodes of the tree may be linked, and that the only admissible choice functions are those that make corresponding choices at linked nodes. (I sketch the details in an appendix.)²⁵ With this addition to the model, the natural comparative similarity relation specified, and the selection functions defined in terms of it, will get the right result, both about *Morgenbesser* cases, and about those which contrast with the *Morgenbesser* cases. And in a plausible model of this kind for Edgington's plane crash case, we will get the result that, even if there is no fact of the matter about

²⁵ Thomason and Gupta considered a *Morgenbesser* case (though not under this name), and the way the choice functions need to be restricted to account for such cases. See Thomason and Gupta (1980: 82–4).

whether Dorothy would have survived if she had made the plane, it is nevertheless true that she probably would have survived (and that, in a sense, it is probably *true* that if she had made the plane, she would not have survived.)

We asked, at the start of Section 7.2, what is the point of giving a truth-conditional form to the conditional assertion account? The answer was that there would perhaps be little point if all conditional assertions were the fragile kind that conform to Adams's thesis, and that reflect only ephemeral features of a local epistemic situation, but that the truth-conditional form facilitated an account of the continuity between these conditionals and more robust ones, including counterfactuals—conditionals that seem to reflect factual judgments about the objective world. But this answer works only if we have reason to give a truth-conditional account of the more robust conditionals that seem to express some kind of items of information that can be detached from the contexts in which they are expressed, and assessed as correct or not, right or wrong. But it is easier to justify a truth-conditional account of the more robust conditionals. It is not only that it seems an intuitive strain to withhold the words 'true' and 'false' in the cases where we judge the counterfactual to be 'right' or 'wrong', 'correct' or 'incorrect'. It is that these items of information (unlike the more ephemeral conditional assertions) combine naturally with each other, and with statements and clauses that uncontroversially have truth conditions. The traditional defense of a truth-conditional approach—that it gets the compositional semantics to work—is an important part of the rationale.

Edgington has used the slogan 'objectivity without truth' to sum up her approach to counterfactuals. I was struck by the juxtaposition of this slogan with the title of a book by Max Kölbel defending relativistic semantics, *Truth without Objectivity*.²⁶ The first slogan is represented by statements like 'She would have been killed if she had made the plane' in the revised version of Edgington's example: the statement was judged to be objectively probable, but neither true nor false. The second slogan fits with a semantics that assigns truth-values to points of evaluation (such as choice functions) that do not represent something there is an objective fact of the matter about. I think we need both objectivity without truth, and truth without objectivity for a satisfactory account of conditionals.

Appendix

I will give a very rough sketch of a simple theory for modeling chance processes. I borrow some resources from game theory to define the models.

A game-theoretic representation of an extensive form game is a tree structure with the branches of the tree representing the choices available to one of the players at the point in the game represented by the node of the tree from which the branches branch. The nodes are labeled by the player who controls the choice at that node. In games involving chance events, one of the 'players' is named 'chance' or 'nature', and for this player's choice points, the branches are assigned probabilities.

In so-called 'games of imperfect information', distinct nodes of the tree may correspond to a single choice point. Suppose, for example, player 2's choice of paper, scissors or rock is

²⁶ Kölbel (2002).

temporally later than player 1's choice of one of these options, but the choices must be made independently—as if simultaneously. In this case, player 2's choice point corresponds to three different nodes of the tree, the ones that result from each of player 1's possible choices. In the game-theoretic representation, the three nodes that follow player 1's choice are said to be *informationally equivalent*, and the set of the three equivalent nodes is an *information set*. The notion of an information set is usually explained intuitively in epistemic terms: player 2 must choose in ignorance of player 1's choice. But the distinction concerns the causal structure of the game, and is not essentially epistemic.²⁷ Player 2's choice is informationally independent of player 1's choice if and only if the choices are causally independent, which implies that if player 1 had chosen differently, player 2's choice would still have been the same. So informational equivalence can apply to chance moves (where epistemic notions don't apply, since the 'player' is not an agent), as well as to moves of rational agents. So suppose we have a game that involves the kind of situation described in the Morgenbesser examples discussed in the text. At the root node of the tree a player chooses whether to bet heads on a coin flip or to decline the bet, leading to two nodes that result from the two possible choices. Then the coin is flipped, independently of whether the bet was accepted. If the player declined the bet, and the coin then lands heads, the player can truly say that if he had accepted the bet, he would have won. The difference between the case where the result of the flip is causally independent of the prior choice and the case where it is not is represented in the model by difference between a case where the two nodes that represent the alternative choices are informationally equivalent and a case where they are not.

A pure chance model is a special case of a game—one that really takes us out of the realm of game theory, since there are no real players, just chance. We have the same kind of tree structure, with information sets in cases of causally independent chance events. Since every branch starts from a node 'controlled' by chance, every branch is assigned a probability value.

Here is a formal definition of this kind of structure.²⁸ (To avoid technical complications, I will restrict attention to finite models.)

A chance model is a structure (H, c) , where H is a finite set of finite sequences meeting these conditions:

- (1) $0 \in H$
- (2) If $(a_1 \dots a_m) \in H$, and $k < m$, then $(a_1 \dots a_k) \in H$.
- (3) If $(a_1 \dots a_m) \in H$ and $i \neq j$, then $a_i \neq a_j$.

This definition determines a finite tree structure. The elements of H may be thought of either as the nodes of the tree (with 0 the root node), or as all the paths through the tree up to a certain point (since there is a one-one correspondence between the nodes of the tree and the paths from the root to that node). The elements of the sequences are the branches of the tree; they represent *outcomes* of the chance event that occurs, or that might occur, at that point in the process. We define Z —the set of *terminal nodes* or *complete histories*—as those $h \in H$ for which there is no a such that $(h, a) \in H$. For each non-terminal node, we can define the branches from that node as follows: $A(h) = \{a : (h, a) \in H\}$.

²⁷ One problem with the epistemic explanation of informational equivalence is this: it is sometimes assumed that players know that all players act rationally, and also know what the payoffs for all the players are. In a game (such as a one-shot prisoners' dilemma) where players make their choices independently, but where each player has a dominating choice, these assumptions will imply that player 2 knows what player 1 did, even though player 2's choice is informationally independent of player 1's choice.

²⁸ See Osborne and Rubinstein (1994: 200) for a definition of an extensive form game with imperfect information that defines the tree structure in this way.

Condition (3) above prohibits an element from occurring more than once on any given path. An element may occur more than once on the tree, but we add the following condition that constrains how this may occur:

- (4) Either $A(h_i) = A(h_j)$ or $A(h_i) \cap A(h_j)$ is empty.

Two nodes h_i and h_j are defined as *informational equivalent* if and only if $A(h_i) = A(h_j)$.

The second element of the structure that defines the model, c , is a function assigning probability values to each branch of the tree. So for any node h , $c(h)$ will be a real number in the interval $(0,1]$. For any nonterminal node h , the values of c for the members of $A(h)$ must sum to one.

In the general case of a model of an extensive form game, one can define *strategies* for each of the players, which are functions taking each of that player's choice points to a choice. A strategy profile is a sequence of strategies, one for each player, which determines not only a complete path through the tree, but also answers to questions about what players would have done if choice points that were not reached had, contrary to fact, been reached. We can define an analogue of a strategy profile for our pure chance model: A *choice function* is a function taking each choice point to one of the choices available at that point. We may also define an *extended choice function* as a function that in addition orders the alternative available at choice points where there are more than two available options.

Now we can take a possible world to be a complete path through the tree, but we might also use a more fine-grained abstract object to be the points of evaluation in a possible worlds model: complete choice functions, or extended choice functions. Our model determines a probability function for the space of possibilities, defined in this way. To determine the weight to be assigned, in the probability space, to a given choice function, just take the product of the probabilities of each choice at each choice point.

Suppose we take choice functions as the points of evaluation in our semantics—the 'possible worlds'. There is then a natural comparative similarity relation that can be used to define a selection function for interpreting conditionals. The idea is simple: first, maximize agreement on all *reached* choice points; second, among choice functions that are tied with respect to this condition, maximize agreement on all choice points. These conditions will not yield a total ordering of all choice functions, relative to each choice function, but it will suffice to determine a unique closest possible world for many potential suppositions, including all where the proposition supposed specifies a particular counterfactual event.

Here are two contrasting examples: (1) there are two urns, one with 9 red balls and 1 black, the other with 9 black and 1 red. First a fair coin is flipped, and if it is heads, a ball is selected at random from the first urn, and if tails from the second urn. The coin landed heads, and a black ball was drawn. What would have happened if the coin had landed tails? Our 'actual' choice function will determine a choice at the unreached choice point where a ball is drawn from the second urn, and this choice will be preserved in the 'closest' choice function to the 'actual' one. (2) Two coins are flipped independently, the first landing heads, the second tails. Since the flips were independent, the result of either would have been the same if the other had been different, and our model gives this result. But suppose our counterfactual supposition is that the two coins land the same way (either both tails or both heads). Which would it have been? Our account gives no answer to this question, even given a full choice function, which seems intuitively right. This is just like Quine's notorious question whether if Bizet and Verdi had been compatriots, they would have been French or Italian. In this kind of case, we not only should not expect a truth-value for the counterfactual, we also should not expect a probability value.

For those cases where the semantics yields a determinate truth-value for a conditional (relative to each choice function), it will also yield a determinate probability value, relative to each node of the model.

It is, of course, a fiction that the choice functions are the possible worlds if this requires that there be a fact of the matter about which choice function is the actual one. But we can say that conditionals (and propositions generally) are true relative to a given history (represented by a path through the tree) if and only if true for all choice functions that determine that path. Some counterfactuals will lack truth-values in the supervaluation, but they still will have well-defined probability values, determined by the chance model, and the path through its tree.

References

- Bennett, J. (2003) *A Philosophical Guide to Conditionals*. Oxford: Clarendon Press.
- Edgington, D. (2004) Counterfactuals and the benefit of hindsight. In P. Dowe and P. Nordhof (eds), *Cause and Chance: Causation in an Indeterministic World*. London and New York: Routledge.
- Gibbard, A. (1980) Two recent theories of conditionals. In W. Harper, R. Stalnaker, and G. Pearce (eds), *Ifs*, Dordrecht: Reidel, pp. 211–47.
- Hájek, A. and Hall, N. (1994) The hypothesis of the conditional construal of conditional probability. In E. Eells and B. Skyrms, *Probability and Conditionals: Belief Revision and Rational Decision*, Cambridge: Cambridge University Press, pp. 75–111.
- Grice, P. (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
- Iatridou, S. (2000) The grammatical ingredients of counterfactuality, *Linguistic Inquiry* 31: 231–70.
- Kaufmann, S. (2004) Conditioning against the grain. *Journal of Philosophical Logic* 33: 583–606.
- Kölbel, M. (2002) *Truth without Objectivity*. London and New York, Routledge.
- Lewis, D. (1973) *Counterfactuals*. Cambridge, MA: Harvard University Press.
- Lewis, D. (1975) A problem about permission. In E. Saarinen et al. (eds), *Essays in Honour of Jaakko Hintikka*, Dordrecht: Reidel, pp. 163–75.
- Lewis, D. (1976) Probabilities of conditionals and conditional probabilities, *Philosophical Review* 95: 581–9.
- MacFarlane, J. (2003) Future contingents and relative truth, *Philosophical Quarterly* 53: 321–36.
- McGee, V. (2000) To tell the truth about conditionals, *Analysis* 60: 107–11.
- Osborne, M. and Rubinstein, A. (1994) *A Course in Game Theory*. Cambridge, MA: The MIT Press.
- Skyrms, B. (1980) *Causal Necessity*. New Haven: Yale University Press.
- Skyrms, B. (1984) *Pragmatism and Empiricism*. New Haven: Yale University Press.
- Stalnaker, R. (1968) A theory of conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, Oxford: Blackwell, pp. 98–112.
- Stalnaker, R. (1975) Indicative conditionals, *Philosophia* 5: 269–86.
- Stalnaker, R. (1984) *Inquiry*. Cambridge, MA: The MIT Press.
- Stalnaker, R. (2011) Conditional propositions and conditional assertions. In A. Egan and B. Weatherson (eds), *Epistemic Modality*, Oxford: Oxford University Press, pp. 227–48.
- Stalnaker, R. (2014). *Context*, Oxford: Oxford University Press.
- Stalnaker, R. (2015) Counterfactuals and Humean reduction. In B. Lower and J. Schaffer (eds.), *A Companion to David Lewis*. Wiley-Blackwell, pp. 411–24.
- Stalnaker, R. and Jeffrey, R. (1994) Conditionals as random variables. In E. Eells and B. Skyrms, *Probability and Conditionals: Belief Revision and Rational Decision*, Cambridge: Cambridge University Press, pp. 31–46.
- Thomason, R. (1970) Indeterminist time and truth-value gaps, *Theoria* 36: 264–84.

- Thomason, R. and Gupta, A. (1980) A theory of conditionals in the context of branching time, *Philosophical Review* 89: 65–90.
- van Fraassen, B. (1966) Singular terms, truth value gaps, and free logic. *Journal of Philosophy* 63: 481–95.
- van Fraassen, B. (1976) Probabilities of conditionals. In W. Harper and C. Hooker, *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science*, vol. 1, Dordrecht: Reidel, pp. 261–308.

Grammar Matters

Sabine Iatridou

It is a great honor to have been asked to contribute to the Festschrift for Dorothy Edgington.* When I was contacted by the editors, my initial reaction was that they had the wrong person and told them so. My work is mostly on the syntax and the syntax–semantics interface, with some morphology occasionally thrown in. The editors claimed they did not have the wrong person. In the end, I hesitatingly accepted their rejection of my self-proclaimed irrelevance and started wondering what I could say that might be of interest to a philosophical audience—of course, more specifically, what I could say about the grammar of conditionals, one of Edgington’s most famous topics, that might prove useful to philosophers. To have any chance of doing this successfully, I would first need to find out what exactly philosophers believe about conditionals and grammar and identify possible misconceptions in those beliefs—because after all, confirming correct beliefs may be less helpful and is definitely less fun. But doing this thoroughly is, of course, an impossible proposition. Even so, I have made an attempt to look for assumptions or explicitly stated beliefs about the grammatical form of conditionals. I will address some of those.

8.1 What’s in a Name?

The first point is one of nomenclature and therefore not ‘deep’. In addition, my impression is that most, if not all, philosophers are well aware of it. Even so, I would still compulsively like to make it.

There is a certain type of conditional that some refer to as ‘counterfactual conditionals’, and that are frequently referred to by philosophers as ‘subjunctive conditionals’. I do not know where this terminology originated, but it is clear that the subjunctive is neither necessary, nor sufficient to create counterfactual conditionals.

The first (and too easy to be interesting) argument for the position that the subjunctive is not necessary to create a ‘counterfactual conditional’ is the fact that there are plenty of languages that do not have a subjunctive at all and still have counterfactual conditionals (e.g. Dutch). But even for languages that have a subjunctive, it can be shown that calling these constructions ‘subjunctive conditionals’ is on the wrong track.

* I am grateful to Kai von Fintel, Roumi Pancheva, Bob Stalnaker, Vina Tsakali, and Lee Walters for comments.

To show that the subjunctive is not necessary for a counterfactual conditional, we will go to French. To show that it is not sufficient, we will go to Icelandic.

French has a subjunctive, which appears under verbs of doubt, for example:¹

1. A: Marie avait un parapluie rouge hier
 Marie had a umbrella red yesterday
 'Marie had a red umbrella yesterday'
- B: Je doute que Marie ait / *²a / *avait
 I doubt that Marie have/SUBJ / have/PRS/IND / have/PST/IND
 un parapluie rouge hier
 a umbrella red yesterday
 'I doubt that Marie had a red umbrella yesterday'

However, the subjunctive is not used in counterfactual conditionals:

2. Si Marie avait / * ait un parapluie rouge, il l'aurait/ *ait
 if Marie have /PST/IND/SUBJ a umbrella red he it have/COND/have/SUBJ
 vu
 seen
 'If Marie had a red umbrella, he would have seen it'

As can be seen above, the subjunctive appears neither in the antecedent, nor in the consequent.

In other words, there are languages that have a subjunctive mood, but do not use it in counterfactual conditionals.

In Iatridou (2000) (see Section 8.2), I argued that in French, as well as in a number of other languages, what is necessary in the morphological make-up of 'counterfactuals' is Past tense (and in some languages, Imperfective Aspect) and the subjunctive appears only if the language has a paradigm for the past subjunctive. As far as I know, no counterexamples to this generalization have been brought forth. Modern French does not have a past subjunctive. Its subjunctive is unmarked for tense. Hence, it cannot appear in 'counterfactual' conditionals. Previous stages of French, however, did have a subjunctive which varied for tense, that is, there was a past subjunctive, and in that stage of the language, the past subjunctive was required in a 'counterfactual' conditional. Modern French, on the other hand, uses the indicative, as it has no past subjunctive.³

¹ Abbreviations:

PRS=present tense

PST=past tense

FUT=future

IND=indicative mood

SUBJ=subjunctive mood

COND=conditional mood

² The star '*' indicates ungrammaticality.

³ In French 'counterfactuals', the antecedent is in the indicative mood, and the consequent is in what grammars refer to as 'le conditionnel' ('Conditional Mood'). However, the latter seems to be nothing more than the combination of past+future in the indicative that we find in other languages, like English and Greek. The only difference between the three languages lies in whether the past and the future are independent words or bound morphemes.

In French we found an argument that the subjunctive is not necessary to form ‘counterfactual’ conditionals. If we look at Icelandic,⁴ we will see that the subjunctive is not sufficient to make a conditional ‘counterfactual’. Icelandic has a past subjunctive and, as predicted by the generalization in the previous paragraph, uses past subjunctive to form ‘counterfactual’ conditionals. However, there is also an environment where the subjunctive appears in a conditional without this being a ‘counterfactual’. Let me first introduce some general background to the phenomenon at large.

In certain languages, English among them, in certain conditional antecedents, the verb can appear in the position where *if* appears, namely, just before the subject:

3. a. If I had known that you were sick, I would have visited you
- b. Had I known that you were sick, I would have visited you

That the verb *had* appears in the position of *if* can be seen from the fact that it necessarily precedes the subject. Many linguists talk about the verb ‘moving’ to the position where the lexical item *if* would otherwise have appeared but for present purposes, I will use the term ‘conditional inversion’. ‘Inversion’ refers to the fact that the positions of the verb and the subject ‘invert’, that is, they exchange places. Inversion can also be seen in matrix questions in English, where the verbs *has* and *is* precede the subject, while they follow it in an assertion:

4. a. **Has** he left already? (compare to *He has left already*)
- b. What **is** she singing? (compare to *She is singing the Marseillaise*)

The term ‘conditional’ in ‘conditional inversion’ obviously refers to the fact that the inversion we are dealing with occurs in conditional sentences. In many languages, English among them, conditional inversion happens only in ‘counterfactual’ conditionals, not in non-counterfactual ones. Contrast (3) to (5):

5. a. If he is sick, I will visit him
- b. *Is he sick, I will visit him

In other languages, Icelandic among them, inversion can also⁵ happen in non-counterfactual conditionals. In non-counterfactual conditionals, when conditional

In English, the future (‘will’) is a separate verb, which can, therefore, carry the past tense. In Greek, the future is an undecidable particle, so the past tense has to go on the matrix verb. In French, both the future and the past tense are bound morphemes, so both have to go on the verb:

- i. English ‘counterfactual’ consequent: Future+Past Verb; e.g. **would leave**
- ii. Greek ‘counterfactual’ consequent: Future Verb+Past: tha efevye
 FUT leave+Past
- iii. French ‘counterfactual’ consequent: Future+Past+Verb; e.g. partirait
 leave+FUT+Past

We are therefore not justified in calling the verb form in the French ‘counterfactual’ consequent a special mood (‘le conditionnel’). The particular form of the verb follows exactly from French being a language like English and Greek, in terms of the semantic needs of the consequent, but differing from them only in the bound versus free status of the morphemes involved.

⁴ Or for that matter, older stages of French.

⁵ To my knowledge, inversion in non-counterfactual conditionals appears in a proper subset of the languages that have inversion in counterfactual conditionals. That is, I do not know of a language that has inversion in non-counterfactual conditionals but does not have it in ‘counterfactual’ conditionals.

inversion does not happen, the verb is in the present indicative, as one would expect (contrast 6a to 6c). But when there is conditional inversion, the verb is necessarily in the present subjunctive (6b versus 6d):⁶

- 6a a. Ef hann hefur farið, kem ég⁷
 if he has/PRES/IND gone, come I
 ‘If he has left, I will come’
- b. Hafi hann farið, kem ég
 has/PRES/SUBJ he gone I come
 ‘if he has left, I will come’
- c. *Ef hann hafi farið...
 if he has/PRES/SUBJ gone
- d. *Hefur hann farið...
 has/PRES/IND he gone

It is unclear what the difference in meaning is between (6a) and (6b), or even if there is one to begin with.⁸ However, it is completely certain that (6b) is not a ‘counterfactual’ conditional.⁹

In other words, what we see here is that the subjunctive in a conditional is not sufficient to make it a ‘counterfactual’.

⁶ Data from Iatridou and Embick (1994).

⁷ The only inversion that matters for us is the one in the antecedent. The inversion in the consequent is irrelevant for our purposes as Icelandic is a ‘Generalized Verb Second language’ and any constituent that is sentence-initial (like the conditional antecedent in our case) must be followed immediately by the verb.

⁸ The only difference that Iatridou and Embick (1994) found between inverted and non-inverted antecedents is that the former cannot be focused. This holds for Icelandic inverted non-counterfactuals as well (thanks to Johannes Johnsson for the data):

- i. Aðeins ef Jón kemur, mun ég fara
 only if John comes will I go
 ‘Only if John comes, will I go’
- ii. Aðeins komi Jón, mun ég fara
 *only comes John, will I go
- iii. A: Undir hvaða kringumstæðum munt þú koma?
 Under what conditions will you come?
 B: Ef Jón fer
 if John leaves
 B’: *Fari Jón
 *leaves John

⁹ Putting aside the question of whether the subjunctive necessarily turns a conditional into a ‘counterfactual’ (it doesn’t), Bjorkman (2011) argues that there is a close relationship between subjunctive and conditional inversion crosslinguistically (not just in Icelandic) and that this can also be seen in English, where *were* is a residue of the subjunctive, which used to be much more productive in earlier stages of the language:

- i. If he were/was absent, the chair would have been offended
 ii. Were/*was he absent, the chair would have been offended

As for the difference between the two expansions in (i), I have found that it is mostly, if not entirely generational. Older speakers prefer *were*, younger ones *was*. It seems that this little residue of the subjunctive of the verb BE is leaving the language.

Since the subjunctive is neither necessary nor sufficient for ‘counterfactual’ conditionals, the term ‘subjunctive conditionals’ is inappropriately used. And as a correlate, the term ‘indicative conditionals’ is inappropriate for non-counterfactual conditionals, as there are plenty of languages where ‘counterfactual’ conditionals are in the indicative mood: those that do not have a past subjunctive, and those that do not have a subjunctive at all.

Throughout this section, I consistently (and almost certainly, annoyingly) kept on putting the word *counterfactual* in quotes. The reason is that while I have been arguing that the term ‘subjunctive’ is not correct, I am not advocating using the term ‘counterfactual’ either. This term is not appropriate either, for at least two reasons. First of all, the constructions in question can be used in contexts where the antecedent and consequent are not contra-to-fact, a fact made famous by at least Anderson (1951) (*But look! If he had the measles, he would have exactly the symptoms he has now*). Moreover, *Future Less Vivids* (FLV), like examples (8b) below, talk about situations that can still be realized and are, therefore, not contra-to-fact. Yet, as will be shown below, FLVs are marked by the same morphological means as ‘counterfactual/subjunctive’ conditionals are. In other words, neither the term ‘subjunctive’ nor the term ‘counterfactual’ is appropriate. What term is then? In co-taught classes and ongoing collaborative work (von Fintel and Iatridou 2019), Kai von Fintel and I have been using the maximally neutral term ‘X-marking’ to refer to that combination of morphemes, that, among other effects it achieves (and we will see some of its other achievements shortly), it turns what philosophers call an ‘indicative’ conditional into a ‘subjunctive’ one. As a mnemonic, one can think of ‘X’ as standing for ‘eXtra’ marking. This would be in opposition to ‘O-marking’, for the morphology that appears in what philosophers call ‘indicative’ conditionals. ‘O’ could be taken as a mnemonic for ‘Open’ or ‘Ordinary’.

So from now on, I will be using the term ‘X-marked’ and ‘O-marked’ conditionals.

8.2 If It Is Not the Subjunctive, then What Is It?

If it is not the subjunctive that yields an X-marked conditional, then what does? After all, it is clearly something about the *form* of the sentence.

There are languages that have a specialized X-marker. One such language is Hungarian, where X-marked conditionals differ from O-marked conditionals in the addition of the marker ‘nV’¹⁰ (Aniko Csirmaz, p.c.). The difference in meaning between (7a,b) and (8a,b) is reflected in the English translations.

7. a. Ha János tudja a választ, Mari (is) tudja a választ
if J knows the answer-acc M (too) knows the answer-acc
‘If John knows the answer, Mary knows the answer’
- b. Ha János tudná a választ, Mari is tudná a választ
if J know.NA the answer-acc Mari too know.NA the answer-acc
If John knew the answer, Mary would know the answer

¹⁰ ‘V’ stands for vowel here. The specific choice of vowel depends on the vowel of the preceding syllable.

8. a. ha holnap el-indul, a jövő hétre oda-ér
 if tomorrow away-leave the following week. onto there-reach
 'If he leaves tomorrow, he will get there next week'
- b. ha holnap el-indulna, a jövő hétre oda-érne
 if tomorrow away-leave.CF the following week. onto there-reach.CF
 'If he left tomorrow, he would get there next week'

However, there are also languages where there is no such thing as a specialized X-marker but whose speakers still clearly know that they are dealing with meanings like (7b, 8b), as opposed to (7a, 8a). How is this possible? In Iatridou (2000), I attempted to explore this question and showed that the morphological means used are pooled from other parts of the grammar. Specifically, in many languages, English among them, there is a past tense morpheme that is not interpreted temporally and in many of these languages there is, in addition, an imperfective morpheme that is not interpreted as an imperfective, though I will not focus on the latter in the current context. I called these morphemes 'fake past' and 'fake imperfective' but one should not read much into the choice of the term 'fake'. I merely meant that the meaning of this morpheme in X-marked conditionals is not what it is in other environments. Let me illustrate.

Consider the pair of sentences in (9), which clearly show that the adverb 'right now' is incompatible with past tense:

9. a. She had a car last year
 b. *She had a car right now

However, in a conditional, the combination *now*+past tense is just fine, yielding a Present Counterfactual¹¹ (PresCF). In other words, the situation described does not hold at the time of utterance:

10. If she had a car right now, she would be happy

Similarly, the adverb *tomorrow* is not compatible with past tense but in a conditional, this combination yields a Future Less Vivid (FLV; see fn 11):

11. a. He left yesterday
 b. *He left tomorrow
 c. If he left tomorrow, he would get there next week

In addition, the presence of a fake past can be detected in sentences that contain a temporally interpreted past morpheme as well, i.e. there is a 'fake' past in addition to a 'real' past, that is, a past tense morpheme that is interpreted temporally (on the fairly common assumption that English pluperfect can be described as containing two instances of past tense). This combination yields a Past Counterfactual (PastCF),

¹¹ Please take the term 'counterfactual' in 'Present Counterfactual' with the grain of salt pointed to in the previous section. What is relevant here, is the Present tense orientation of the conditional. Idem for the term 'Past Counterfactual', that will come up soon.

which indicates that the situation described does not hold at a time before the time of utterance:

12. a. He was descended from Napoleon
- b. *He had been descended from Napoleon
- c. If he had been descended from Napoleon he would have been shorter

It seems that in all these cases, the fake past morpheme is somehow involved in yielding the relevant part of the meaning in which X-marked conditionals differ from O-marked ones. The actual temporal interpretation of the conditional is what it would have been without this fake past. Specifically, the conditional in (10) is interpreted as a PrsCF because without the fake past, its temporal interpretation would be about the present:

13. 'If she had a car right now' – fake past = if she has a car right now

The conditional in (11c) is interpreted as an FLV because without the fake past, its temporal interpretation would be about the future:

14. 'If he left' – fake past = if he leaves

Finally, the conditional in (12c) is a PstCF because without the fake past, its temporal interpretation would be about the past:

15. 'If he had been descended from Napoleon' – fake past = If he was descended from Napoleon

In other words, X-marked conditionals receive the temporal interpretation of the corresponding O-marked conditionals.¹²

In (2000) I suggested one way how this might be done. I proposed a meaning for the past tense morpheme that is neither that of temporal past, nor that of contra-to-factness (again, with the relevant grain of salt regarding Anderson cases and FLVs). This basic meaning turns into that of temporal Past or contra-to-factness after the addition of elements from the environment.¹³

Since then there have been other proposals in the literature about how fake tense does what it does. The reader can consult the original paper for details, as well as subsequent work by others (e.g. Ana Arregui, Michela Ippolito, John Mckay, Katrin Schulz) that aim to improve on that proposal.

X-marking in fake past languages does not just consist of fake past, however. As I described in 2000, X-marking in many languages also includes a fake imperfective. That is, an imperfective morpheme that is not interpreted (necessarily) as an

¹² This may have bearings on a debate that I understand exists in philosophical circles, namely whether X-marked conditionals are very different from O-marked conditionals, including in basic properties. The default conclusion from the discussion in the main text would be that X-marked conditionals differ from their O-marked counterparts only in what X-marking contributes. My understanding is that Edgington (1995) explores the possibility that counterfactuals are past tense indicatives.

¹³ Under this proposal, there is no 'fake past', obviously. This adjective was used descriptively to refer to non-Past uses of the 'Past' morpheme. In my (2000) proposal, there is no morpheme which unambiguously means 'Past'.

imperfective. In other words, the task is to find out what X-marking contributes, that is, how X-marked conditionals differ in their meaning from O-marked conditionals, and then investigate how this meaning is compositionally contributed by the fake past and fake imperfective morphemes.

There is a crucial piece of the puzzle that one has to keep in mind when taking on the aforementioned task, and I would like to lay this out before closing this section. X-marking does not only have the discussed effect in conditionals. It has at least two other roles.

In Iatridou (2000), I argued that in many languages, X-marking is used to construct what are (misleadingly¹⁴) called ‘counterfactual wishes’:

16. She wishes she was taller than she is

English has a lexicalized verb ‘wish’, but in many languages, the way to say ‘wish’ is to take the verb ‘want’ and add X-marking to it. In fact, I had argued for a certain relation between the X-marking that appears in conditionals and the X-marking that appears in wishes. Specifically, I had argued that there is a crosslinguistic tendency to use consequent X-marking on the verb ‘want’, and antecedent X-marking on the embedded verb:

17. The Conditional/Desire generalization:

- a. X-marked conditional: if p_{m1} , q_{m2}
- b. X-marked desire: I want_{m2} that p_{m1}

One clear example is Spanish, where antecedent X-marking consists of Past subjunctive, and consequent X-marking of the fake past+future combination we saw earlier, which is often referred to as the ‘Conditional Mood’ in the literature on Romance languages.

Spanish X-marked conditional:

18. Si **fuera** más alto **sería** un jugador de baloncesto.
 If be.3.sg.PST.SUBJ more tall be.3.sg.COND a player of basketball
 ‘If s/he was taller, s/he would be a basketball player’

Spanish X-marked desire:

19. **Quisiera** que **fuera** más alto de lo que es.
 Want.3.sg.COND that s/he be.3.sg.PAST.SUBJ more tall than it s/he is
 ‘I wish s/he was taller than s/he is’

A third part of grammar where X-marking is at play, is the construction of the weak necessity modal ‘ought’ (von Stechow and Iatridou, 2008):

20. You ought to do the dishes but you don’t have to

¹⁴ Misleadingly, because the desire itself is an actual world desire. That is (16) reflects the subjects preference structures in the actual world. What is ‘counterfactual’ is the situation described in the embedded clause.

Again, as with wishes, English has a lexicalized item but in many languages, the way to construct the weak necessity modal is to take the strong necessity modal and add X-marking to it. For example, Spanish, again:

21. #Debo limpiar los platos, pero no estoy obligado
must clean the dishes but not am obliged
#‘I must do the dishes but I am not required to’
22. Deberia limpiar los platos, pero no estoy obligado
must.COND clean the dishes but not am obliged
‘I must do the dishes but I am not required to’

In other words, when one sets out to explore how X-marked conditionals are constructed, one should also keep in mind that whatever one says about X-marking when studying conditionals, this very same X should be able to turn *want* into *wish*, and *must* into *ought*.¹⁵

8.3 The Mark of *Then*

Even a cursory perusal of the literature, shows that conditionals are referred to interchangeably as *if p, q* and *if p, then q*. However, the switch from one form to the other is not innocent. In this section we will see differences between them that make this point.¹⁶

For many cases, the effect of *then* seems negligible:

23. a. If Pete runs for President, the Republicans will lose
b. If Pete runs for President, then the Republicans will lose

But for several other cases, *then* seems impossible:

24. a. If I may be frank (*then) you are not looking good today
b. If John is dead or alive (*then) Bill will find him
c. If he were the last man on earth (*then) she wouldn’t marry him
d. Even if you give me a million dollars (*then) I will not sell you my piano

The difference between (23b) versus the sentences in (24), is that the latter all intend to assert the consequent. (24a) is a ‘relevance conditional’,¹⁷ a type of conditional in which the *if*-clause does not contain the conditions in which the consequent is true but in which it is relevant. In (24b) the *if*-clause is such that it exhausts all possibilities, hence the consequent is asserted. In (24c), the *if*-clause is chosen in such a way as to make a conversational move in which the consequent is asserted. Similarly for (24d).

A rough approximation, in other words, of the contribution of *then* is that it brings with it a presupposition:

¹⁵ Some of the challenges involved in this task are discussed in von Stechow and Iatridou (2019).

¹⁶ The discussion is based on Iatridou (1994). See Hegarty (1996) for an improvement. See Izvorski (1997) for a generalization of the proposal to other correlative pro-forms.

¹⁷ Also sometimes known by the name ‘biscuit conditionals’.

- 25. a. Statement: if p , then q
- b. Assertion: if p , q
- c. Presupposition: there are some $\sim p$ cases that are $\sim q$

It is obvious that the sentences in (24) violate the presupposition in (25c), as they leave no room for the existence of $\sim p \& \sim q$ cases.

On the other hand, this is not the case for (23b), where the presence of *then* contributes something like (25), cast within possible-world talk:

- 26. In some possible worlds epistemically accessible to me in which Pete does not run for President, the Republicans win.

When we force the acceptability of *then*, we force the existence of $[\sim p, \sim q]$ cases. For example, what would otherwise have been a relevance conditional, becomes something that Mary Poppins might have said, who was able to turn a situation of one being hungry into a situation in which a sandwich magically appears in the fridge:

- 27. If you get hungry then there will be a sandwich in the fridge

And in (28), we are forced to consider cases that do not fall under 'rainy' or 'sunny'. That is, 'rainy' and 'sunny' together should not exhaust all possible weather conditions, if we want *then* to be acceptable:

- 28. If it is rainy or sunny then I will visit you (but if it is foggy, I will not)

Without *then*, (28) could have been taken to convey that I will visit you no matter what. But with *then* we are forced to take $\sim p$ possibilities into account. This is not possible at all in some cases, like (24b), where the existence of $\sim p$ cannot be accommodated.

Finally, we can see the effect of *then* when the antecedent is a presupposition of the consequent. In such a case, the $\sim p$ cases that are crucial to the presupposition brought in by *then*, will make the consequent suffer from presupposition failure. Consider the following sentences:

- 29. a. If [Peter smiles at her]_i Kathy likes it_i
- b. If Peter cooks [something]_i, he gives half of it_i to Kathy

As they are, the sentences in (29) are fine but once we introduce *then* they become variably¹⁸ odd, because if Peter does not smile or cook something, the pronoun *it* in the consequent will suffer from existential presupposition failure.

In short, *if p, q* cannot be used interchangeably with *if p, then q*.¹⁹

¹⁸ I say 'variably' because my impression is that speakers need a little more time to compute the oddity that results from inserting *then* in (29) compared to (24). I suspect that parsing the correct reference of the pronoun might take a bit, but this is only an intuition. For example, Roumi Pancheva (p.c.) suggests that in (29b) there may be a parsing strategy in which the anaphora might be some form of modal subordination – 'the thing that he would have cooked'.

¹⁹ *then* is also impossible in *only if* conditionals:

- i. Only if it rains will we stay inside
- ii. *Only if it rains then we will stay inside
- iii. *Only if it rains then will we stay inside

8.4 There Is No Magic in *if*

The item ‘*if*’ is often used as short for ‘conditionals’. However, the presence of the item *if* is not necessary to have a conditional interpretation. For one, we can have what we called above ‘inversion’, where the verb appears in the position of *if*:

30. a. If I had known you were sick I would have visited you
 b. Had I known you were sick I would have visited you

Even though in English, conditional inversion is restricted to X-marked conditionals, in other languages, it can also take place in O-marked conditionals. Above we saw Icelandic being such a language. German is as well:

31. a. [Wenn Hans kommt] geht Susanne
 if Hans comes goes Susanne
 ‘If Hans comes, Susan goes’
 b. [Kommt Hans] geht Susanne
 comes Hans goes Susanne
 ‘If Hans comes, Susan goes’

While inversion can happen in a number of environments, including questions, as we saw in (4), inversion of a tensed²⁰ verb in an adjunct can *only* receive a conditional interpretation (Iatridou and Embick, 1993). This generalization holds crosslinguistically; at least no counterexamples have been reported so far. In other words, a sentence like (23b) can never, for example, mean ‘Because I had known, ...’ and (24b) can never mean ‘Because Hans comes, Susan will leave.’

This means that (30b) is just as much ‘necessarily’ a conditional as (30a), and (31b) is just as much necessarily a conditional as (31a) is, even though *if* is missing in both (30b) and (31b). And by ‘necessarily’ I mean that the grammatical form of all four sentences *only* permits a conditional interpretation. If (30b) and (31b) do not contain *if*, yet receive a conditional interpretation, why then do we consider that *if* is the *sine qua non* of conditionality? It clearly is not.²¹

At first blush, the presupposition in (8c) would predict that *then* should be perfect with *only if p*, *q* as in the latter ALL $\sim p$ cases are $\sim q$ cases. However, discussing this will take us too deeply into syntax, which does not seem appropriate in the current context. For more details on the effect of *then* on *only if p, q* please see Iatridou (1994).

²⁰ It is crucial that the verb be tensed for a conditional interpretation. If we have inversion with a participle for example, the meaning is completely different. Consider Italian:

- i. avendo Gianni finito il giornale, iniziò a leggere il libro
 having Gianni finished the newspaper, started to read the book
 ‘Gianni having finished the newspaper, he started reading the book’

²¹ I would dare venture the following guess, in fact: it may well be the case that the verb can move to the position of *if* and kick it out so to speak, exactly because *if* has no meaning of its own. Items like *because* or *although* can never be replaced by a verb because they do have a meaning of their own, which would not be recoverable under deletion. But if *if* does not contribute conditionality, how do we know to interpret sentences like (30a,b) and (31a,b) as conditionals? The answer may lie in the tense and aspect composition of the verbs, as well the construction as a whole. As we will see shortly, there is good reason to believe that all we need from the syntax to access a conditional interpretation is information of which clause to interpret as the restrictor and which as the scope of a quantifier over worlds.

The absence of *if* in the above sentence is the result of conditional inversion. There are quite a few grammatical constraints on conditional inversion and inversion in general and syntacticians have successfully explored and explained many of them. For example:

- 32. a. If I knew the answer, I would tell you
b. *Knew I the answer, I would tell you
- 33. a. He knew the answer
b. Did he know the answer?
c. *Knew he the answer?
d. *Did I know the answer, I would tell you
- 34. a. Had he not seen the truck?
b. Hadn't he seen the truck?
c. Had he not seen the truck, he would have been killed
d. *Hadn't he seen the truck, he would have been killed

This is not the appropriate place to delve deeper into the syntax of inversion; the interested philosopher is encouraged to look up his or her friendly neighborhood syntactician and ask about 'T-to-C movement'. The syntactician will understand this term and will know what to say.

There are also semantic and pragmatic effects of conditional inversion.

An inverted antecedent cannot be focused (see also n. 7). This generalization holds for all the languages in which it has been tested.²² For example, it cannot be a fragment answer:

- 35. A: When/under what conditions would Mary have come?
B: If she had been offered many artichokes
B': *Had she been offered many artichokes

Nor can it be focused by *only*:

- 36. a. Only if you had given me a million dollars would I have sold you my piano
b. *Only had you given me a million dollars would I have sold you my piano

Nor can it be focused in sentences called 'clefts':

- 37. a. It is if Walter had come that Susan would have left
b. *It is had Walter come that Susan would have left

So we learn two basic things from inverted conditionals: The item *if* is not necessary to form a conditional²³ and furthermore, different morphosyntactic expressions of

²² Data here are from Iatridou and Embick (1993). See also Horn (2000) and Biezma (2011).

²³ One could also make the quick and easy argument that *if* is not sufficient either, as this item appears in embedded questions as well (not just in English, in many other languages as well):

- i. I do not know if he will be able to get here on time

This argument is a bit too easy, though, because the historical origins of this homophony are unclear.

conditionality come with their own slew of interpretive properties. Grammatical form matters, in other words. But they are still all conditionals.

In note 21, I suggested that maybe the reason that *if* can be absent in conditionals is that it does not contribute to the interpretation of the sentence. One might wonder why, if *if* has no meaning, it is there to begin with. In syntax, there are conditions on the wellformedness of sentences as such. In fact, syntax is full of them. Often, these conditions take the form of the need for words that do not contribute to the semantics. One easy to spot example is the appearance of dummy verb *do* in non-subject questions:

- 38. a. What did you eat?
- b. *What you ate?
- c. When did he leave?
- d. *When he left?

but

- 39. Who ate the tiramisu?

The item *if* is called a complementizer. Complementizers are words that introduce clauses. The item *that* is a complementizer in the following example:

- 40. He thinks **that** she never calls him

Stowell (1981) found that in English, among other languages, complementizers may be optional when the clause they introduce is the object of a verb, as in (33), but they are required when the clause they introduce is not the object of a verb. That is, the complementizer in (40) can go missing:

- 41. He thinks she never calls him

But the complementizer cannot go missing when the clause is in subject position:

- 42. a. that she never calls him bothers him
- b. *she never calls him bothers him

Similarly, a conditional antecedent is a clause and specifically, a clause that is not in the object position of a verb. It is what is called an 'adjunct'. Therefore, its complementizer cannot go missing:²⁴

- 43. a. We will go to the movies if it rains
- b. *We will go to the movies it rains

If this path of thinking is correct, the presence of *if* is dictated by syntactic reasons and not because it makes a particular semantic contribution.

I will conclude this section by mentioning that some languages may not even have an item like *if*. By this I mean that they do not have a morpheme that marks an adjunct clause as an antecedent of a conditional as such, yet, they have no problem

²⁴ In the case of inversion, the verb moves to the position of the complementizer, as we said. So even though the lexical item *if* is missing, the complementizer *position* is filled.

expressing conditionals. This may be the case for Turkish, in fact, as I argued in Iatridou (2013).

8.5 And, There Is No Special Status to *if p, q*

In the previous section I showed that one does not need *if* to make a conditional. In this section I will show that one does not need *if p, q* either. That is, *if p, q* is a particular syntactic form that *leads* to a conditional semantics. It is wrong to consider *conditionality* coextensive with the form *if p, q*.

We actually have conditionals with forms that are even farther removed from the old and familiar *if p, q* than the sentences with inversion like the ones we have seen so far. For example, take a look at this conjunction (Culicover and Jackendoff, 1999):

- 44. a. She looks at him and he shies away in fear²⁵
 b. = if she looks at him, he shies away in fear

Moreover, the two conjuncts do not even have to be propositions. The first conjunct can be a nominal or an imperative:

- 45. a. One more mistake and you are fired
 b. =if you make one more mistake, you will be fired
- 46. a. Ignore your homework and you will fail
 b. = if you ignore your homework you will fail

Lest the reader doubt that ‘Ignore your homework’ is, in fact, an imperative (because after all English morphology is quite poor and that form could be just about anything) we can go to languages where the imperative is explicitly marked as such, and we will see that we are definitely dealing with an imperative:²⁶

- 47. agnoise ta mathimata su ke tha kopis (Greek)
 ignore/IMPER the lessons yours and FUT cut
 ‘Ignore your lessons and you will fail’

The sentences in (44a, 45a, 46a, and 47) clearly receive a conditional interpretation. Therefore, why should we not call them conditionals? The only reason why somebody might not do that is because s/he thinks that ‘conditional’ is the name for a particular *morpho-syntactic* form, namely the one that has an adjunct clause introduced by *if*, which is also the syntactic form chosen for the paraphrases in (44b) and (44b). But I find it hard to believe that when philosophers talk about ‘conditionals’ that they think they are referring to a particular *syntactic* construction. I assume they think they are referring to a particular *interpretation*. But if that is the case, (44a, 45a, 46a, and 47) have to be included in this class as well.

²⁵ Culicover and Jackendoff show that certain tense and aspect combinations are required for a conditional interpretation of (44a) and that those also hold for (44b):

- i. She has looked at him and he has shied away in fear (≠ conditional)
- ii. If she has looked at him, he has shied away in fear (only epistemic)

²⁶ See Kaufmann (2011) and von Stechow and Iatridou (2017) and references therein for more on this use of the imperative.

Again, we see that *if* or the syntax associated with it does not have a privileged status when it comes to conditionality. And like before, we can also see that the choice of grammatical form determines possible interpretive choices. For example, conjunctions of this sort cannot yield epistemic conditionals (48), and interpretations associated with X-marking are restricted (49b) ((49)a is meant to show that X-marking on the consequent is in principle acceptable):

- 48. His light is on and he is at home
 ≠ If his light is on, he is at home
- 49. a. One more mistake and he would have been fired
 b. *She had looked at him and he would have shied away in fear
 ≠ If she had looked at him, he would have shied away in fear

Moreover, conjunctions like the one in (46a), that is, with an imperative first conjunct have certain restrictions on the predicates involved (Bolinger, 1967):

- 50a. Own a piece of property in this town and you get taxed mercilessly
 = If you own a piece of property in this town, you get taxed mercilessly
- 51a. Own this property and I'll buy it from you
 ≠ If you own this property, I will buy it from you

In short, we have the same bifurcated conclusion: Any sentence form that receives a conditional interpretation has to be classified as a conditional and studied as such. But grammatical form matters, as not all forms that receive a conditional interpretation have the same type of restrictions. There are quite a few more syntactic constructions that these two points can be made with, but I will mention only one more. The following sentence has been argued to receive a conditional interpretation (Stump 1985) but its form is obviously very different from *if p, q*:

- 52. a. Standing on a chair²⁷, he will be able to reach the ceiling
 b. = If he stands on a chair, he will be able to reach the ceiling

Sentence (44a) clearly receives a conditional interpretation but if we change the predicate slightly, the meaning immediately shifts:

- 53. a. Having long arms, he will be able to reach the ceiling
 b. ≠ If he has long arms he will be able to reach the ceiling
 c. = Because he has long arms, he can reach the ceiling

I hope the general point has come across by now: if we study only conditionals that have the syntactic form *if p, q* we narrow our vision considerably. We need to study a variety of different grammatical forms that receive a conditional interpretation. This way we will also be able to understand why and how and which possible meanings group together for each grammatical expression of conditionality.

To be honest, the mistake of identifying the interpretive category 'conditional' with the *syntactic* form *if p, q* is also committed by linguists. Culicover and Jackendoff

²⁷ Note that this is not a case of conditional inversion. There is no subject and the verb is not tensed. It is a participle.

(1999) explore sentences like (44) and claim that they have identified what they call a 'syntax-semantics mismatch'. They argue that this particular type of *and* is syntactically a coordination (conjunction) but in the semantics, the sentence receives a conditional interpretation and this is a case of 'subordination' according to them. This poses a challenge to an approach they are arguing against: there is a common belief in generative grammar according to which semantic interpretation is read off of a level of syntactic representation, called 'Logical Form' or LF. The surface string, that is, the sentence that we pronounce, yields an (unpronounced) LF via a series of syntactic operations. Cases like (44) are a problem for an LF-based approach, according to Culicover and Jackendoff, because there are no syntactic transformations that will change a coordination into a subordination. This much is indeed true, there are no syntactic operations that we know of that will transform a structure of coordination into a structure of subordination. But do we need such an operation?

When Culicover and Jackendoff use (44a) to argue against LF-based approaches, they assume that the latter would need to turn the **syntax** of coordination into the **syntax** of *if p, q*, which indeed is a case of syntactic subordination. But the **syntax** of *if p, q* is not the same as 'semantic subordination' or 'conditional semantics'. It is merely one of the syntactic structures that can receive a conditional interpretation. There is in principle no need for a syntactic construction to first turn into a different syntactic construction in order to map onto a particular semantic scheme. In order to prove a syntax-semantics mismatch, they would need to give a semantics for conditionals for the semantic side of the 'mismatch' and show that it is not possible to map into that from a particular syntactic construction. But they do not do this. What they do instead, is give syntactic structures for both sides of the alleged mismatch. This is because they wrongly identify the essence of *conditional semantics* with the *syntactic structure if p, q*. But one should not. The syntactic structure *if p, q* is one of several syntactic structures that can yield conditional semantics, as we saw. And it is not the case that those other syntactic constructions should first turn into the *syntactic construction if p, q* before they receive a conditional interpretation. Why would they need to?

To prove a mismatch, one would need to first assume a certain *semantics* of conditionals, and show that it cannot be derived compositionally from a certain syntax. But they do not assume any conditional semantics. As mentioned above, they identify a particular syntactic construction with a conditional semantics.

So let us assume Kratzer's semantics for conditionals, which is currently one of the most popular theories for conditionals in the linguistic community. According to Kratzer, in conditionals, one clause restricts a modal/quantifier over worlds (what we call 'the antecedent') and another clause is the scope (what we call 'the consequent'). This means that what we need from the syntax is an indication as to which clause is the restrictor and which clause is the scope. One such indication can be seen in the syntax of *if p, q*: the adjunct (whether it has the item *if* or not) is the restrictor. But why should that be the only possible morpho-syntactic flag? Why should there be only one syntactic construction with instructions for which clause is the restrictor and which the scope of the quantifier over worlds? We have another indication with *and* in (44): the first conjunct maps into the restrictor and the second conjunct into

the scope. The order of the conjuncts cannot change with the conditional interpretation remaining intact.

54. a. She looks at him and he shies away in fear
b. $\neq$ He shies away in fear and she looks at him

So in the conjunction in (44a, 54a) the order of conjuncts tells us what the syntax to semantics mapping should be. In the syntactic construction *if p, q*, the cues are different: there it is the adjunct that maps into the restrictor and the matrix into the scope. Given those cues, the order of the two can flip, unlike in (54), where the order itself was the cue to the semantic mapping. With *if p, q*, we do not rely on word order to decide which clause is the antecedent and which the consequent:

55. a. If she looks at him, he shies away in fear
b. = He shies away in fear if she looks at him

And we should also contrast (54a) with (56), a garden variety conjunction, where the two clauses can be switched without effect on the meaning:

56. a. London is the capital of England and Paris is the capital of France
b. = Paris is the capital of France and London is the capital of England.

The inability to flip the two conjuncts (54a) is exactly because we would then lose the grammatical cue as to which clause is the antecedent.

In short, coordinations like (44) and *if p, q* structures contain the same amount of information that a conditional semantics needs, at least for the identification of the restrictor and of the scope of the modal. There is no one privileged syntactic structure of a conditional semantics that all the other ones would have to first turn into, before being mappable to a conditional semantics.

At this point, I would like to preempt a possible thought in the reader's mind, which if it is there, is the result of a bias and has no grounds. The reader might think 'All that is fine and well but the form *if p, q* is really what conditionals are and coordinations like (44) are marginal structures.'

There is no basis for such a belief, however. Coordinations of this sort are cross-linguistically extremely widespread (von Stechow and Iatridou 2017). They contain all the syntactic information one would want for a conditional semantics and are immediately and very easily identified as such by speakers. They, and other constructions like the ones we mentioned, are conditionals just as much as those of the *if p, q* form.

8.6 Conclusion

In honor of Dorothy Edgington, I have tried to provide a gentle introduction to a grammarian's view of conditionals for philosophers.

I zigzagged through an assortment of grammatical properties of conditionals, with one of my main goals having been to show that grammatical form matters: different syntactic expressions of conditionality come with a different range of possible meanings. Moreover, I argued that we should not consider 'conditionals' coextensive

with the syntactic form *if p, q*. The syntactic construction *if p, q* is merely one of several syntactic paths that lead to a conditional semantics. I hope this point is relevant because I assume that when philosophers talk about ‘conditionals’, they are talking about a particular *interpretation*, not a particular syntactic form. Overly narrowing conditional semantics to only one syntactic construction makes it harder to identify where each of the elements of meaning originates.

References

- Anderson, A. (1951) A note on subjunctive and counterfactual conditionals. *Analysis* 12: 35–8.
- Beck, S. (1997) On the semantics of comparative conditionals. *Linguistics and Philosophy* 20(3): 229–71.
- Bhatt, R. (2003) Locality in correlatives. *Natural Language and Linguistic Theory* 210: 485–541.
- Biezma, M. (2011) Conditional inversion and GIVENNESS. In *Proceedings of SALT* 21: 552–71.
- Bjorkman, B. (2011) BE-ing default: The Morphosyntax of Auxiliary Verbs. MIT PhD Thesis.
- Bolinger, D. (1967) The imperative in English. In Halle, M., H. Lunt, H. McClean, and C. van Schooneveld (eds) *To Honor Roman Jakobson*, vol. I. The Hague: Mouton. pp. 335–62.
- Culicover, P. and Jackendoff, R. (1999) Semantic subordination despite syntactic coordination. *Linguistic Inquiry* 28(2): 195–217.
- Edgington, D. (1995) On Conditionals. *Mind* 104: 235–329.
- Fintel, K. von and Iatridou, S. (2008) How to say *ought* in foreign: The composition of weak necessity modals. In J. Guéron and J. Lecarme (eds), *Time and Modality*. New York: Springer.
- Fintel, K. von and Iatridou, S. (2017) A modest proposal for the meaning of imperatives. In A. Arregui, M. Rivero, and A. Salanova (eds), *Modality across Categories*. Oxford: Oxford University Press
- Fintel, K. von and Iatridou, S. (ms. MIT). Prolegomena to a theory of X-marking.
- Hegarty, M. 1996. The role of categorization in the contribution of conditional *then*: Comments on Iatridou. *Natural Language Semantics* 4: 111–19.
- Horn, L. (2000) Pick a theory (not just any theory). In L. Horn and Y. Kato (eds), *Negation and Polarity*, pp. 147–92. New York: Oxford University Press.
- Iatridou, S. (1994) On the contribution of conditional *then*. *Natural Language Semantics* 2: 171–99.
- Iatridou, S. (2000) The grammatical ingredients of counterfactuality. *Linguistic Inquiry* 31(2): 231–70.
- Iatridou, S. (2013) Looking for Free Relatives in Turkish (and the unexpected places this leads to). In *Proceedings of WAFL* 8. Umut Özge ed. MITWPL, Cambridge MA.
- Iatridou, S. and Embick, D. (1994) Conditional inversion. In *Proceedings of North East Linguistic Society* 24th. UMass, Amherst: GLSA, pp. 189–203.
- Izvorski, R. (1997) The syntax and semantics of correlative proforms. In K. Kusumoto (ed.), *Proceedings of the Twenty Sixth Annual Meeting of the North Eastern Linguistic Society, Amherst, Mass.: GLSA*; pp. 133–47.
- Kaufmann, M. (2011) *Interpreting Imperatives*. Dordrecht, Heidelberg, New York: Springer.

- Kratzer, A. (2012) *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.
- Stowell, T. (1981) Complementizers and the empty category principle. In V. Burke and Pustejovsky, J. (eds), *Proceedings of the Eleventh Annual Meeting of the North Eastern Linguistic Society, Amherst, Mass.: GLSA*; pp. 345–63.
- Stump, G. (1985) *The Semantic Variability of Absolute Constructions*. Dordrecht: Reidel.

Constructing the Impossible

Kit Fine

A number of philosophers have flirted with the idea of impossible worlds and some have even become enamored of it. But I think it is fair to say that it has not met with the same degree of acceptance as the more familiar idea of a possible world. Whereas possible worlds have played a broad role in specifying the semantics for natural language and for a wide range of formal languages, impossible worlds have had a much more limited role; and there has not even been general agreement as to how a reasonable theory of impossible worlds is to be developed or applied.¹

The reasons for this relative neglect are somewhat hard to pin down. But one principal reason, I suspect, has to do with the fact that impossible worlds are not theoretically robust; they lack the theoretical virtues that we expect of a framework within which to conduct semantical investigation. Their theoretical shortcomings have perhaps three related sources. The first is that we would like our semantics to be compositional; and not only that—we would like the compositional clauses for the logical connectives to be ‘uniform’ or non-disjunctive. This is a theoretical virtue in itself but, without uniformity, it is not even clear that we will have clauses for the logical connectives themselves as opposed to some gerry-mandered product of the theoretician’s mind.

But as a number of philosophers have noted, this requirement leads to disaster once impossible worlds are added to the mix. When only possible worlds are in question, the clause for negation is:

(*) the statement $\neg A$ is true in a world iff it is not the case that A is true in the world.

But if there are impossible worlds—or, at least, logically impossible worlds—then presumably there is an impossible world in which both a statement A and its negation $\neg A$ are true. So then A is true in the world and, by the above clause, it is not the case that A is true in the world. A contradiction.

¹ I have presented this chapter as a talk at a number of different venues. They include the philosophy departments at University of Vermont, Buffalo University, Washington University and the University of Toronto, and a metaphysics workshop in Tucson; and I should like to thank the audiences at those venues and also Lee Walters for helpful comments.

It is with admiration and affection that I dedicate this chapter to Dorothy Edgington, who has made so many singular and important contributions to philosophical logic; and I am especially heartened that in a number of her writings (1985, 2010), she has endorsed the present approach to modality in terms of possible situations rather than possible worlds.

The second issue concerns the range of impossible worlds. One of the key roles of impossible worlds is to provide us with a more refined conception of entailment. Under the possible worlds approach, certain given statements will entail another if any possible world in which the given statements are true is one in which the other statement is true. Once impossible worlds are allowed, we obtain a more refined conception of entailment: certain given statements will entail another if any possible or impossible world in which the given statements are true is one in which the other statement is true. So when only possible worlds are in question, a figure being trilateral may well entail its being triangular, while this may not be so when impossible worlds are also taken into account.

With the more refined conception of entailment comes a more refined conception of equivalence, or mutual entailment; and with the more refined conception of equivalence comes a more fine-grained conception of proposition, under which the proposition expressed by a statement may essentially be identified with the set of possible or impossible worlds in which it is true. Thus statements which express the same proposition under the possible worlds approach may be true in different impossible worlds and hence end up expressing different propositions once those other worlds are taken into account.

But even though we may wish to restrict the notion of entailment or to adopt a more fine-grained conception of proposition, it is not likely that we will wish to jettison all non-trivial entailments or all non-trivial identities between propositions. Thus even though we may wish to distinguish between the belief in different mathematical truths, it is not clear, in general that we will wish to distinguish between the belief that $P \ \& \ Q$ and the belief that $Q \ \& \ P$; and, likewise, even though we might wish to distinguish between the counterfactual supposition of different mathematical falsehoods, it is not clear, in general, that we would wish to distinguish between the counterfactual supposition that $P \ \& \ Q$ and the counterfactual supposition that $Q \ \& \ P$.

But under an 'anarchic' policy in which any form of impossible world is allowed, all of these different beliefs and counterfactual suppositions would be distinguished and it is not even clear that impossible worlds would serve to make distinctions among these different propositions that might not be made by other, more straightforward, means. What we need, in order to account for some moderate form of rationality in our beliefs or some moderate form of discriminability in our counterfactual suppositions, is some intermediate position in which some impossible worlds are allowed and others not.

How then is the range of the impossible worlds to be restricted? Which of the putatively impossible worlds is a genuine impossible world? A natural thought, in answering this question, is that what impossible worlds there are should somehow be a function of what possible worlds there are, that whatever constraints there are on the genuinely possible worlds should somehow serve as constraints on the genuinely impossible worlds.

But how is the possible to serve as a constraint on the impossible? The answer that most naturally suggests itself is that the impossible worlds should respect the entailments induced by the possible worlds, that if certain given statements are true in an impossible world then some other statement should be true in that

world as long as it is true in any possible world in which the given statements are true.

However, such a requirement means that every statement whatever will be true in any impossible world and hence that the impossible worlds can do no work in demarcating a more refined notion of entailment or a more fine-grained conception of proposition. For in any impossible world, there will be statements that are true but not true in any possible world. But these statements will then vacuously entail any statement whatever (with respect to the possible worlds); and so any statement whatever will be true in the impossible world.

The third issue has to do with the application of impossible worlds to linguistic or intuitive data. Suppose one wants to explain why the separate necessity of A and of B implies the necessity of $(A \ \& \ B)$ while the separate possibility of A and of B does not imply the possibility of $(A \ \& \ B)$. The possible worlds semantics for the modal operators provides a ready explanation. For if A and B are separately true in all possible worlds (as required by the semantic clause for necessity) then so is $(A \ \& \ B)$, while if A and B are separately true in some possible world (as required by the semantic clause for possibility) then it does not follow that $(A \ \& \ B)$ is true in some possible world. And we see the same pattern of explanation repeated in other contexts.

But consider now some analogous data to which the apparatus of impossible worlds might be applied. How is it that in saying $(A \ \wedge \ B)$, one thereby says A (and also says B) while, in saying A, one does not thereby say $(A \ \vee \ B)$?² Using the apparatus of impossible worlds, one might say that in any impossible world in which $(A \ \wedge \ B)$ is true A (and also B) will be true, while A might be true in an impossible world without $(A \ \vee \ B)$ being true in that world. If one says C just in case C is true in all of the 'assertively' accessible worlds (possible or not), then one sees how in saying $(A \ \wedge \ B)$ one says A, since A will be true in all of the accessible worlds in which $(A \ \wedge \ B)$ is true, while in saying A one may not be saying $(A \ \vee \ B)$, since $(A \ \vee \ B)$ may fail to be true in a world in which A is true.

But this is not much of an explanation. The puzzling difference between ' $\wedge$ ' and ' $\vee$ ' has simply been reduplicated at the level of the semantics; and one often has the feeling, when appeal is made to the apparatus of impossible worlds, that no genuine explanatory purpose has been served and that we merely have an ad hoc reduplication of the phenomenon to be explained.

I do not claim that these reasons are decisive. But anyone with a feeling for theoretical virtue should surely be disturbed by the difficulty of incorporating impossible worlds into the framework of possible worlds in a natural and seamless way.

In recent years, I have been working on a version of situation semantics - one might call it 'truthmaker semantics'—which is meant to provide an alternative to possible worlds semantics. One of the things that has struck me about this alternative semantics is how easily it is able to accommodate the impossible. Rather than being an artificial addition to the possibilist semantics, the impossible emerges as a natural -

² The data are perhaps even clearer in the case of telling someone what do to do. In telling someone to X & Y, I am thereby telling them in part to X (and to Y) but, in telling them to X, I am not thereby telling them (even in part) to X or Y.

one might almost say *inevitable*—extension of the possible, in much the same way in which the system of real numbers emerges as a natural extension of the rational number system or the system of complex numbers emerges as a natural extension of the real number system. It is the aim of this paper to show how this is so; and, if I am successful, then this will constitute an argument for the admission of the impossible into semantics—something which I myself have been slow to appreciate—but also for truthmaker semantics itself as a viable and valuable alternative to the possible worlds approach.

I begin with an exposition of a standard approach to truthmaker semantics, using possible states in place of possible worlds (Section 9.1). I go on to describe a key construction, analogous to the extension of the rationals to the reals, for extending a space of possible states to one that also contains impossible states (Section 9.2). This has a number of advantages—mathematically and in theory and application—over the more usual approaches (Section 9.3–9.4). I then describe another construction, somewhat analogous to the extension of the reals to the complex numbers, which provides further resources for countenancing the impossible and further applications (Section 9.5).

I conclude with a lengthy formal appendix. The proofs are, for the most part, straightforward and will have a familiar feel to someone already acquainted with lattice theory and the theory of partial order. It should be possible to read the informal exposition and the formal appendix independently of one another, but each will make more sense if read in conjunction with the other.

The knowledgeable reader will have noticed that I have said next to nothing about some of the issues that have been most prominent in the recent literature on impossible worlds.³ Philosophers have been intrigued by the ontological status of impossible worlds. Do they exist and, if they do exist, then do they have the same status as possible worlds? To my own mind, these questions are of peripheral interest. The central question is whether impossible worlds or the like are of any use, especially for the purposes of semantic enquiry. If they are of no use, then who cares whether they exist or what they are like? And if they are of some use, then we should be able to find a place for them within our ontology, if only as a convenient fiction.

9.1 Truthmaker Semantics with Possible States

Underlying the present approach is a *state space*, or domain of states. This plays the same role vis-à-vis the truthmaker semantics as the pluriverse of worlds plays vis-à-vis the possible worlds semantics. Initially, we shall think of the state space as comprising possible (and actual) states. Thus it may comprise the state of this patch being red all over as well as the state of its being blue all over, but it will not include the state, if there be such, of the patch being both red and blue all over.

Possible states may, of course, be partial in a way in which possible worlds are not. Thus the state of the patch being red all over has no bearing on whether it is raining in Timbuktu, although any possible world will either include or exclude its raining in Timbuktu. Possible states, in contrast to possible worlds, also enjoy a

³ Berto (2009) reviews some of the recent literature on the topic.

mereological structure: one possible state may be a proper part of another. Thus the state of the patch being red and round will contain, as a proper part, the state of its being red.

A state space, then, is a domain of states endowed with mereological structure. We may say that a given state is an *upper bound* of some others if the others are all part of the given state; and we may say that a given state is the *sum* or *fusion* of some others if it is the least upper bound of the others, i.e. if it is an upper bound and a part (proper or improper) of any upper bound. If the states of a state space are meant to be possible states then there is no reason, in general, why the fusion of arbitrary states should exist. There will, for example, be no fusion of the states of the patch being red all over and of its being blue all over since such a state, were it to exist, would be an impossible state. However, it is plausible to suppose that some arbitrary states will have a fusion as long as they have an upper bound. For if there is some state having the other states as parts, then those other states will be jointly compatible and so there will be nothing to prevent them from having a fusion.

We may say that a space is *complete* if any states within the space have a fusion and that the space is *bounded complete* if any states within the space will have a fusion whenever they have an upper bound. The upshot of the present discussion, when stated in these terms, is that a space of possible states can be assumed to be bounded complete but cannot, in general, be assumed to be complete simpliciter.

Let us now show how to develop a semantics for sentential logic against the background of a state space.⁴ The semantics will tell us which states within a state space will verify or falsify a given statement. There are various senses in which a state may be said to verify or falsify a given statement but, in considering the clauses below, it is important to bear in mind that our intention is that the state should exactly verify or falsify the given statement, i.e. that the state should be *wholly* relevant, and not just relevant in part, to the truth or falsity of the statement.

Here then are the clauses for negative, conjunctive and disjunctive statements:

- (i) (a) A state verifies a negative statement $\neg A$ just in case it falsifies the unnegated statement A , and
 (b) it falsifies the negative statement $\neg A$ just in case it verifies the unnegated statement A ;
- (ii) (a) A state verifies a conjunction $(A \wedge B)$ just in case it is the fusion of a state that verifies A and a state that verifies B , and
 (b) it falsifies the conjunction $(A \wedge B)$ just in case it falsifies A or falsifies B ; and
- (iii) (a) A state verifies a disjunction $(A \vee B)$ just in case it verifies one of the disjuncts A or B , and
 (b) it falsifies the disjunction $(A \vee B)$ just in case it is the fusion of a state that falsifies A and a state that falsifies B .

⁴ A semantics of this sort was first proposed by van Fraassen (1969).

There are a number of variants on these clauses that might be considered. It might be allowed, for example, that a verifier for $(A \wedge B)$ should also be a verifier for $(A \vee B)$ (and, likewise, that a falsifier of $(A \vee B)$ should also be a falsifier for $(A \wedge B)$). The semantics might also be extended to the quantifiers. But the present approach to impossible states will be largely independent of how exactly these variants or extensions of the semantics are given.

It is important to note that the fusions specified on the right of clauses (ii)(a) and (iii)(b) may not exist. Thus even though there is a possible state s that verifies A and a possible state t that verifies B , there may be no fusion of the states s and t and hence no state that verifies the conjunction $(A \wedge B)$. In particular, given that there is no verifier for A compatible with a falsifier for A , no state will verify $(A \wedge \neg A)$.

It should also be noted that the present semantics will provide us with a much more refined conception of entailment even without the admission of impossible states. For let us take a statement to *entail* another if, as a matter of logic, any verifier for the one is a verifier for the other. Then A will entail $(A \vee B)$, since any verifier of A is, ipso facto, a verifier of $(A \vee B)$. But $(A \wedge B)$ will not in general entail A , since there is in general no reason to suppose that a verifier for $(A \wedge B)$ (which includes a verifier for B) will be wholly relevant to the truth of A .

9.2 The Construction

We now wish to extend the previous semantics to one which accommodates impossible states. There are two aspects of the extension—one purely ontological and the other purely semantical. On the ontological side, we wish to extend the space of possible states to one that also includes impossible states without regard to how these states might be put to semantical use. On the semantical side, we wish to state semantic clauses over an extended space of possible and impossible states without regard to how the extension was made.

We begin with the ontological question. Suppose we are working within a space of possible states. Then certain fusions of those states—of the states of a patch being red and of its being blue, for example—may not exist. A very natural approach to the idea of an impossible state is that it should correspond to the fusion of two or more possible states when their fusion would not otherwise exist. It is, so to speak, a fictitious fusion of states, which is to behave in exactly the same as the fusion *were* it to exist.

An impossible state, on this view, might be represented by a set of possible states, whose fusion it is meant to be. But if impossible states are meant to behave in the same way as fusions, then not every distinct set of possible states can be taken to correspond to a distinct fusion. One problem is this. Let s be the possible state of a patch being red and round and let t be the possible state of the patch being blue. Then the set $\{s, t\}$ will correspond to an impossible state—one which, intuitively, is the fusion of s and t . Now let s' be the possible state of the patch being red. Then the set $\{s, s', t\}$ will also correspond to an impossible state. But, intuitively, this impossible state will be the same as the impossible state corresponding to the set $\{s, t\}$; for, given that s' is a part of s , the fusion of s, s' , and t should be the same as the fusion of s and t .

There is another problem, which operates, so to speak, in the opposite direction. Suppose now that s'' is the state of the previous patch being round. Then the set $\{s', s'', t\}$ will correspond to an impossible state. But, intuitively, this impossible state will be the same as the impossible state corresponding to the set $\{s, t\}$ or, for that matter, to the set $\{s, s', s'', t\}$; for, given that s is the fusion of s' and s'' , the fusion of s', s'' and t should be the same as the fusion of s and t (or of s, s', s'' and t).

This suggests that of all the sets of possible states which correspond, via a fictitious form of fusion, to the same impossible state, we should pick out one as the true 'representative' of the impossible state.⁵ But which of the various sets of possible states should we pick?

It turns out that the two problems above are essentially the only problems that can arise in identifying impossible states with sets of possible states. The first problem can be solved by requiring the identifying sets to be closed under part-whole:

Downward Closure If one possible state is a part of another, then it should be a member of an identifying set if the other is.

The second problem can be solved by requiring the identifying sets to be closed under fusions when they exist:

Upward Closure If one possible state is a fusion of others, then it should be a member of an identifying set if the others are.

Thus we can solve both problems by requiring the identifying sets to conform to Downward and Upward Closure. In effect, we take the representative set to be the largest of the sets corresponding to the given impossible state.

We might call a set of possible states conforming to these two conditions an *ideal*.⁶ Some of the ideals will contain a 'principal' member, of which all the other members are parts. These will correspond, under the construction, to the possible state that is their principal member. The other ideals will contain no such state; and these will correspond to the impossible states or 'virtual' fusions of possible states.

Once given this construction, we may then, with a good conscience, simply postulate that the required fusions will exist. And once given the extension of the state space, either through construction or postulation, the corresponding extension of the semantics for the connectives will be completely straightforward; for we may use exactly the same clauses as before. The only difference will be in their application. Where before the fusion of verifiers for the conjuncts of a conjunction or the fusion of falsifiers for the disjuncts of a disjunction might not exist, their existence is now assured. So, for example, we can now be sure that there is a verifier for $(A \wedge \neg A)$ as long as there is a verifier and a falsifier for A , since the verifier for $(A \wedge \neg A)$ can simply be taken to be their fusion.

⁵ Alternatively, though more clumsily, we might identify the impossible state with an 'equivalence class' of sets of possible states.

⁶ Or, more properly, a *virtual ideal* to distinguish it from the related notion of ideal familiar from lattice theory and the theory of partial order.

9.3 Advantages of the Construction

Quite apart from any other considerations, it is worth emphasizing how natural the present construction is from the mathematical point of view. For the given space of possible states may not be complete, some fusions of states from within the space may not exist; and so there is a natural mathematical need to ‘complete’ the space, i.e. to extend it to a space in which the fusions do exist. And given this need, there is a natural construction, familiar in general outline from other mathematical contexts, by which the need might be met. For we may identify the completing states with ‘fictitious’ fusions of possible states, while imposing minimal constraints on how those fusions are mereologically related.

Of course, there are a number of different ways in which the initial state space might be completed. Instead of adding many different impossible states, for example, we might add a single impossible state, which is to be the fusion of any possible states that would not otherwise have a fusion. The fusion of a patch being red and its being blue, say, would then be the same as the fusion of a figure being a triangle and its being a square. But the present construction is optimal in the sense that it maximizes the number of the impossible states subject to the requirement that any difference in impossible states should be attributable to a difference in the possible states of which they are composed. Further, it is unique in this respect. If we wish to maximize the number of impossible states and if we also want to be able to individuate impossible states in terms of the possible states of which they are composed, then there is no choice but to accept something essentially like the present construction.

We can bring out the naturalness of the present construction by means of an analogy with the reals. The gaps left in the arithmetical ordering of the rational numbers are to be ‘filled’ by the introduction of the irrational numbers and in such a way that the same general arithmetical principles will hold. Similarly, gaps left in the mereological ordering of the possible states are to be ‘filled’ by the introduction of impossible states and in such way that the same general mereological principles will hold. In both cases, the new objects—be they irrational numbers or impossible states—are to be identified in terms of their relationship—be it arithmetical or mereological—to the pre-existing objects, the rational numbers in the one case and the possible states in the other.⁷

In addition to these mathematical virtues, the present construction is able to avoid the difficulties that beset the admission of impossible worlds. In the first place, the clauses for negation—or for the connectives in general—do not lead to contradiction when applied to impossible states. A statement can be both verified and falsified by an impossible state; and since for a statement to be falsified is not for it to fail to be verified, no contradiction ensues. Of course, just as we have countenanced separate clauses for verification and falsification under the truthmaker semantics, we might countenance separate clauses for truth and falsehood under the possible worlds semantics. Instead of the usual clause for negation ((*) above), we might have:

⁷ I hasten to add that the analogy is not exact since the system of rationals is not bounded complete.

- (*)(a) the statement $\neg A$ is true in a world iff A is false in the world,
- (*)(b) the statement $\neg A$ is false in a world iff A is true in the world

and similarly for the other connectives.⁸ Possible worlds will have the special feature that a statement is true in a world just in case it is not false. But once we extend the clauses to impossible worlds, which lack this feature, then no contradiction will ensue from the supposition that a statement is both true and false in a world.

However, from the present perspective, the clauses (*)(a) & (b) constitute an ad hoc departure from the classical semantics as it is usually stated. For the fact is that negation is subject to the stronger condition (*); and so uniformity requires that this stronger condition should also hold for the impossible worlds. The clauses under the truthmaker semantics, by contrast, already require us to distinguish between the verification and falsification of a statement and to state separate clauses for each;⁹ and so there is no shift in how these clauses are to be stated as we move from possible to impossible states.

Another issue concerns the range of the impossible. We wanted the constraints on what is possible to function as constraints on what is impossible, so that not every impossibility whatever was allowed. But within the possible worlds framework it was hard see how this was to be done. In extending the pluriverse of possible worlds into the outer darkness of the impossible, there seemed to be no reasonable constraints on what should or should not be allowed.

This difficulty immediately disappears under the truthmaker approach. For the impossible states are ideals; and the mereological conditions on ideals will severely restrict their make-up and guarantee that they are subject to very much the same behavior as the possible states. Suppose, for example, that an impossible state contains the states of block a being on top of block b and of block b being on top of block c . By Upward Closure, it will contain the state of a being on top of b and b being on top of c ; and by Downward Closure, it will contain the state of a being on top of c , given that this state is a part of the state of a being on top of b and b being on top of c . Thus if an impossible state contains a verifier for the statement ' a is on b ' and also a verifier for the statement ' b is on c ', then it will contain a verifier for ' a is on c ', just as one would expect.

Again, something similar, though less satisfactory, can be done under the possible worlds approach. For if we can fuse possible states then why should it not be possible to fuse possible worlds?¹⁰ One relatively minor problem with this proposal is that it will give us too few impossible worlds—we cannot simply add a single conflicting state to a possible world, for example—and it is not easily amended to give us just the worlds we want. But a more serious worry is that it is unable to predict the right behavior for the resulting fusions. Suppose, for example, that a is on top of b and b on

⁸ This is, in effect, the semantics for first degree entailment within the system R of relevance logic.

⁹ Indeed, two statements with the same possible states as verifiers may not have the same possible states as falsifiers and so it will not in general be possible to treat the falsifiers of a statement as a function of the verifiers.

¹⁰ This—at least in certain respects—is the approach of Brandom and Rescher (1980) and of Restall (1997). Indeed, we might take our state space just to be a space of possible worlds and our construction will then give us the fusions of those worlds. I should note that the original modeling of van Fraassen (1969) might also be obtained as a special case of our construction.

top of c in world w_1 and that c is on top of a and a on top of b in world w_2 . Then we would like b to be on top of a in the ‘fusion’ $w_1 \sqcup w_2$ of the two worlds.

In w_1	In w_2	In $w_1 \sqcup w_2$
a	c	a
b	a	b
c	b	c
		a

But how are we to secure this result given that b is not on top of a in either of the two worlds?

It is not that impossible worlds need to be jettisoned. Within the theory of state spaces, there is no obstacle either to having possible states that correspond to possible worlds or to having even bigger states that correspond to impossible worlds. The point, rather, is that we cannot properly understand how the impossible worlds should behave unless we regard them as being composed of something smaller than the possible worlds. We need to decompose the possible worlds into their parts before we can see how they might sensibly be reconstituted as impossible worlds.

Thus from the present point of view, the problem with the standard approach to impossible worlds lies not in its ontology but in its starting point. It is only by starting with states and seeing worlds as a special case of states that we are able to see how a reasonable theory of impossible worlds might then be developed.¹¹

9.4 Applications

We turn to the all-important question of application. One of the most striking features of the present construction, in addition to its mathematical naturalness, is the ease and versatility with which it can be applied. In what follows, I will focus not so much on the advantages of truthmakers over possible worlds but on the additional advantages that accrue to the truthmaker approach upon the addition of impossible states. I will also not focus on the obvious advantages that derive from having a more refined conception of content but on cases in which the impossible

¹¹ Let us define a *world-state* as a state w with the property that every state is either a part of w or is incompatible with a possible state that is a part of w . A world-state, in this sense, may or may not itself be a possible state. Oddly enough, it can be shown that the original state space of possible states need not contain any world-states even though the extension of the space is bound to contain a world-state (viz., the fusion of all possible states). Thus it may only be by admitting impossible worlds that there are any worlds at all; and in this case, of course, the impossible worlds cannot be obtained via the possible worlds but only via the possible states.

I should also mention—on the other side, so to speak—that if one starts with a pluriverse of possible worlds, then one might identify a possible state with a non-empty set of possible worlds and take the part-whole relation on states to be the relation of set-theoretic containment on the corresponding sets of possible worlds. One might then apply the present construction to obtain an account of impossible states within the standard possible worlds approach although, of course, many distinctions that could be made within the more general state-based approach would then disappear.

states can be put to real and perhaps unexpected explanatory work. However, even within this narrow focus, I have only been able to touch upon a small sample of the possible applications.

Modality It is not normally thought that the semantical analysis of modality requires appeal to impossible worlds. For, after all, a statement may be taken to be possible if it is true in some possible world and to be necessary if it is true in all possible worlds. Impossible worlds do not come into it.

However, informal talk of possibility—or of what might be—is strangely at odds with its formal treatment. If I say that Pete might be in London or Paris, then this seems to imply that he might be in London and that he might be in Paris. Thus it looks as if, under the ordinary use of ‘might’, we are committed to the following rule of inference:

$$\frac{\Diamond(A \vee B)}{\Diamond A}$$

Let me not consider the question of whether we are actually so committed, but just assume that we are.¹²

Then the question arises as to how we might provide a semantics for $\Diamond$ -statements that sanctions the above inference but does not validate the corresponding, obviously invalid, inference for conjunction:

$$\frac{\Diamond A}{\Diamond(A \wedge B)}$$

Within the framework of the truthmaker semantics, an answer immediately suggests itself. For we may take the statement $\Diamond A$ to be true if all of the exact verifiers for A are possible in some appropriate sense of ‘possible’ (perhaps tighter than the sense in which the possible states are taken to be possible). The first of the inferences will then be valid, since any verifier of A is a verifier of $(A \vee B)$ and hence A will be possible if $(A \vee B)$ is possible. However, the second of the inferences will not be valid, since any verifier of A may be possible even though a verifier of $(A \wedge B)$, which may contain extraneous material, is not possible.

However, if the verifiers are restricted to possible states, there will be no verifiers of $(A \wedge \neg A)$ and so $\Diamond(A \wedge \neg A)$ will be ‘degenerately’ valid under the current semantics. Clearly, an undesirable result!

This further problem is readily solved if we admit impossible states, since $(A \wedge \neg A)$ will then have a verifier which is not in fact possible. Thus even for the ordinary modalities there is some point in allowing impossible states.

Partial Content A number of philosophers have thought that there is a notion of partial content, which is some sort of refinement of the more usual notion of logical consequence. Thus the content of A will in general be taken to be part of the content of $(A \wedge B)$ and yet $(A \vee B)$ will not in general be taken to be part of the content of A ,

¹² The analogous question for counterfactuals is discussed in Fine (2012a, 2012b) and the analogous question for permission has been hotly debated in the linguistics and philosophy literature.

even though $(A \vee B)$ is a logical consequence of A and A a logical consequence of $(A \wedge B)$. Our intuitions concerning partial content can be tested via in our intuitions concerning partial truth. I am, in fact, a British philosopher. The statement that I am an American philosopher is then partly true in virtue of the fact that I am a philosopher but the statement that I am American is not partly true in virtue of the fact that I am American or British; and this is because we have a true partial content in the one case but not in the other.

How is the notion of partial content to be defined? Let me here give a natural definition of the notion within the context of truthmaker semantics.¹³ We take the content of a statement to be given by the set of its verifiers and we then define C to be *part of the content of* A if (i) every state in the content of A contains a state in the content of C and (ii) every state in the content of C is contained in a state in the content of A . Thus when A is true, a verifier of A will exist and contain a verifier of C , and so A will be true in part because its partial content C is true; and, when C is true, a verifier of C will exist and be contained in a verifier of A , and so A will be partly true because C is true—just as one would expect.

This account immediately solves the problem over disjunction. For the content of $(A \vee C)$ will not in general be part of the content of A , since a verifier for C (and hence for $(A \vee C)$) may not be contained in a verifier for A . However, there is still a problem over conjunction. For a verifier for A may not be a verifier for $(A \wedge C)$, given that no verifier for C is compatible with the verifier for A . In particular, when C is itself of the form $\neg A$, there will be no verifier for $(A \wedge \neg A)$ and so neither A nor $\neg A$ will be part of the content of $(A \wedge \neg A)$.

Again, this difficulty disappears once we countenance impossible states (and also take every statement to have a verifier). For given a verifier s for A , there will be a verifier t for C , and so the fusion of s and t will be a verifier for $(A \wedge C)$ (possibly impossible) of which s is a part. Thus a reasonable account of partial content seems to require that we admit impossible states as verifiers.

I believe that the notion of partial content has numerous applications in linguistics and philosophy. One, already alluded to, is to the question of defining partial truth or verisimilitude. For we can take a statement to be partially true to the extent that its partial content is true and we can take one statement to be closer to the truth than another to the extent that more of its partial content is true and less of its partial content is false. Another application is to the logic of belief under the assumption of moderate rationality. For a natural suggestion is that, in believing $A_1, A_2, \dots, A_n$ one believes C , just in case the content of C is part of the content of $(A_1 \wedge A_2 \wedge \dots \wedge A_n)$. From this it will follow that in believing $(A \wedge B)$ one will believe A and believe B , even though in believing A one need not believe $(A \vee B)$. Other applications are to the theory of confirmation, the concept of subject matter, and the logic of deontic and imperative statements.

¹³ See Fine (2014). Gemes (1994, 1997) and Yablo (2014) have developed related accounts of partial content. I develop my own account and its relationship to Gemes' and Yablo's in Fine (2017a).

Of special interest is the use of impossible states to represent a given subject-matter. Suppose, for example, that our interest is in the color of a particular patch. This subject-matter can be represented by the various possible states that bear upon it—the patch's being red, green, blue etc. But these states can in their turn be represented by the single impossible state that is their fusion; and we can thereby talk of subject-matters in the same way—and as a special case—of the way in which we talk of states. Thus the combination of two subject-matters, considered as states, will simply be their fusion and their common part will simply be the fusion of their common parts.

Counterfactuals^{14,15} One of the central applications that the proponents of impossible worlds have wished to make is to counterfactuals with counter-possible, and not merely counter-factual, antecedents. For it looks as if we need to distinguish between counterfactual consequences of different counter-possible suppositions. Thus it seems true to say that if Hobbes had squared the circle then he would have squared the circle even though it is not true to say that if Hobbes had found a counter-example to Fermat's Last Theorem then he would have squared the circle. The introduction of impossible worlds provides a way of dealing with such differences. For we may then suppose that there are closest (impossible) worlds in which Hobbes squares the circle and that in those worlds he squares the circle, of course, but does not provide a counter-example to Fermat's Last Theorem.

I have no objection in principle to the use of impossible worlds within this context. But there is a general difficulty in determining what should be true under a counter-possible supposition, which the vague reference to close impossible worlds does nothing to allay. Consider, for example, the counterfactual 'if Hobbes had square the circle, his contemporaries would have been amazed by his mathematical ability' and let us suppose that his contemporaries were not, in fact, amazed by his mathematical ability. Then why do we take the closest impossible world in which Hobbes squares the circle to be one in which his contemporaries were amazed for, after all, we appear to get a closer world if it is one in which, like the actual world, his contemporaries were not amazed. It might be argued in response that it is some sort of law that if Hobbes squares the circle in the circumstances of the time then his contemporaries will be amazed and that we want the truth of this 'law' to be preserved in the closest world. That may be so. But recall, this closest world is an impossible world; and in such an impossible worlds we can have it true that Hobbes squares the circle and that it is a law that if he squares the circle then his contemporaries will be amazed, yet not have it true that his contemporaries will be amazed. Thus when the possible no longer acts as a constraint on closeness, it is no longer clear that an account of counterfactuals in terms of closeness will deliver the intuitively correct results.

¹⁴ Some of these applications are considered in Yablo's book 'Aboutness', Princeton University Press, 2014, though without appeal to the proposed technology of impossible states.

¹⁵ A comparison between Yablo's approach and my own is made in 'Yablo on subject Matter', *Philosophical Studies* 177 (1):129–171 (2020).

My own account of counterfactuals within the truthmaker framework is able to make some headway with this problem. On this account Fine (2012a, 2012b), the counterfactual from A to C is taken to be true if any outcome of a verifier for A will contain a verifier for C . If verifiers are required to be possible states, then a counterfactual with a counter-possible antecedent will be vacuously true, just as with the possible worlds account. But if we allow the verifiers of the antecedent to be impossible states, then there is the possibility of distinguishing between counterfactual statements with different counter-possible antecedents.

This requires that we make sense of the outcomes of an impossible state. But how is this to be done? Given the mereological structure of states, we can make a start on the problem. For let us suppose that the impossible state s can be ‘factored’ into the possible states $s_1, s_2, \dots$ in the sense that (i) s is the fusion $s_1, s_2, \dots$ and (ii) no one of the states $s_1, s_2, \dots$ is a proper part of a possible state that is a part of s . Now each of the respective states $s_1, s_2, \dots$ may have the respective states $t_1, t_2, \dots$ as possible outcomes; and we may then take the fusions of the states $t_1, t_2, \dots$ to be the possible outcomes of s . Thus suppose that my boss tells me to catch flight 105 to Buffalo and also tells me (in a fit of absent-mindedness) to catch flight 106 to Detroit. Then I can correctly say that if I had done what my boss told me to do then I would now be in both Buffalo and Detroit.¹⁶ This result will be predicted on the above semantics since an outcome of my catching flight 105 is that I arrive in Buffalo and an outcome of my catching flight 106 is that I arrive in Detroit.

However, this account only works for certain cases. It is not even clear that it works for the Hobbes case mentioned above (perhaps some kind of coarse-graining of the possibilities would be required in such a case); and considerably more would need to be done to provide a more general account if, indeed, a general account can be given.¹⁷

9.5 Modal Completion

I have suggested one way of introducing impossible states into a state space. But I do not wish to suggest that it is the only way. Indeed, it seems to me that different applications may require different forms of impossibility and that the impossible states countenanced by our construction need not be regarded as the last word on what impossible states there are (just as the extension of the number system to the reals should not be regarded as the last word on what numbers there are).

¹⁶ I might mention, incidentally, that this is a case in which it is appropriate to consider the impossible scenario in which I catch both flights rather than the possible scenario in which my boss says something different and so it provides an especially clear case in which, under the closest world analysis of counterfactuals, an impossible world verifying the antecedent of a counterfactual should be taken to be closer to the actual world than any possible world that verifies the antecedent.

¹⁷ Let me make two general remarks in this connection. First, when counterpossibles are in question, we should probably not admit the rule $A > C / A > C'$ where C' is a classical logical consequence of C but the kind of weaker rule considered in Fine (2012b). Second, Dorothy, in ‘On Conditionals’ (p. 265), has expressed the view that ‘confidence in the counterfactual expresses the judgement that it was probable that B given A , at a time when A had non-zero possibility’. This would seem to rule out counterpossibles, but it is possible that the kind of state spaces we have been considering will make it possible to assign a non-zero probability to impossible antecedents.

Of course, any further extension of this kind means that we will have to tolerate impossible states that cannot be individuated in terms of their possible parts, since our construction serves to generate all such states. Under certain plausible assumptions, the failure of impossible states to be individuable in terms of their possible parts will require that there should be modal 'monsters', impossible states whose impossibility is not attributable to any conflict between their possible parts. One has a natural aversion to such monsters but, all the same, there may be reasons for wanting to allow them and various natural means by which they might be generated.

One possibility, still at a high level of abstraction, concerns the 'opposite' of necessary states. We may say that a state is *necessary* if it is compatible with every (possible) state. Within a state space, there is bound to be at least one necessary state. For any state whatever will vacuously be an upper bound for the null set of states; and so, as long as the state space contains at least one state, the null set of states will have a fusion, which we may designate 'the null state'. But the null state will be necessary, i.e. compatible with any possible state s , since its fusion with s will be s itself.

However, the null state may not be the only necessary state within a state space (and this is a respect in which we cannot simply identify possible states with sets of possible worlds). It might be thought, for example, that, for distinct objects a and b , the state of a 's being identical to a is distinct from the state of b 's being identical to b or that, for any object a , a 's being of a certain kind (or being of a certain kind if it exists) is distinct from a 's being of some other kind, even if a is of both kinds.

Given a necessary state, one might want to countenance an opposite necessary state. In some cases this might correspond to the 'negation' of the necessary state. But not always, since we will want the opposite of a fusion of necessary states to be the fusion of their opposites. Thus the opposite of the state of a 's being identical to a and b 's being identical to b would be the fusion of the state of a 's being distinct from a and b 's being distinct from b , and the opposite of the null state would be the null state itself.

When a state space is supplemented with opposites, we might represent the resulting states by ordered pairs of the form (s, t) , where s is a possible state and t is a necessary state. Intuitively, (s, t) is the fusion of s with the opposite of t . Thus we end up with something like the representation of a complex number as an ordered pair (a, b) of reals, with the first component in (s, t) representing the 'real' or truly possible part of the state and the second component representing the 'imaginary' or impossible part of the state.

One great advantage of the present construction is this. For certain purposes, it is helpful to be able to assume that every statement has at least one verifier. For example, on the account of partial content given above, it will only in general be correct to suppose that the content of A is part of the content of $A \wedge C$ if it can be supposed that C has a verifier. Now given a complete state and adopting the previous clauses for the verification and falsification of complex statements, every complex statement will have a verifier if each atomic statement has both a verifier and a falsifier (not compatible, of course). This requirement, in its turn, will automatically be satisfied for contingent atomic statements but it will not be satisfied for atomic statements that are either necessarily true or necessarily false if we require the verifiers or falsifiers to be possible

states; for no possible state can verify a necessary falsehood or falsify a necessary truth. However, let us suppose that each necessary atomic truth A is verified by at least one necessary state s . Then we can take a falsifier for A to be the opposite of s ; and, likewise, when the necessary atomic falsehood A is falsified by a necessary state.

Further discrimination within the impossible may be desired. Thus we may want to distinguish different ways in which a may not be identical to a , perhaps through being identical to b or to $c \dots$, or different ways for an object not to be of given kind, perhaps through being of this kind or of that kind. But these further discriminations require that we consider the detailed content of the various states and the general forms of construction that we have so far considered will not be adequate to generate them. One must dive into the belly of the beast to know what further impossibilities it will deliver up.

Formal Appendix

Preliminaries

Recall that $\sqsubseteq$ is a *partial order* (po) on S if it is a reflexive, transitive and anti-symmetric relation on S . Given a po $\sqsubseteq$ on S , we shall make use of the following standard definitions (with $s, t, u \in S$ and $T \subseteq S$):

s is an *upper bound* of T if $t \sqsubseteq s$ for each $t \in T$;

s is a *least upper bound* (lub) of T if s is an upper bound of T and $s \sqsubseteq s'$ for any upper bound s' of T ;

s is *null* if $s \sqsubseteq s'$ for each $s' \in S$ and otherwise is *non-null*;

$s \sqsubset t$ (s is a *proper part* of t) if $s \sqsubseteq t$ but not $t \sqsubseteq s$;

s *overlaps* t if for some non-null u , $u \sqsubseteq s$ and $u \sqsubseteq t$;

s is *disjoint from* t if s does not overlap t .

The least upper bound of $T \subseteq S$ if it exists is unique (since if s and s' are least upper bounds, then $s \sqsubseteq s'$ and $s' \sqsubseteq s$ and so, by anti-symmetry, $s = s'$). We denote it by $\sqcup T$ and call it the *fusion* of T (or of the members of T). When $T = \{t_1, t_2, \dots\}$, we shall sometimes write $\sqcup T$ more perspicuously as $t_1 \sqcup t_2 \sqcup \dots$.

A *state space*—which we may also call a *P-space*—is a pair $(S, \sqsubseteq)$, where S (*possible states*) is a non-empty set and $\sqsubseteq$ a relation on S subject to the following two conditions:

Partial Order (PO) $\sqsubseteq$ is a po on S ;

Bounded Completeness (BC) Any subset of S with an upper bound has a least upper bound.

A state space $S = (S, \sqsubseteq)$ is said to be *complete* if every subset of S has an upper bound (and hence, by BC, a least upper bound).

An *extended space*—which we also call an *E-space*—is an ordered triple $(S, P, \sqsubseteq)$, where $(S, \sqsubseteq)$ is a complete state space and P (*possible states*) is a non-empty subset of S subject to:

Downward Closure $t \in P$ whenever $s \in P$ and $t \sqsubseteq s$ (any part of a possible state is also a possible state).

We also say that a state s in an E-space $(S, P, \sqsubseteq)$ is *consistent* if $s \in P$ and *inconsistent* otherwise. The more accurate terms are *possible* and *impossible*, but these terms are subject to an unfortunate ambiguity, since a possible K , where K is a kind of state, may either be a K -state that is possible or a state that is possibly a K -state. A set of states within an E-space $(S, P, \sqsubseteq)$ is said to be *compatible* if their fusion belongs to P and otherwise to be *incompatible*.

In any P-space $(P, \sqsubseteq)$ or E-space $(S, P, \sqsubseteq)$, there will exist a least state $\sqcup \emptyset$, designated by $\wedge$, which will be a part of every state; and in any E-space $(S, P, \sqsubseteq)$, there will also exist a greatest state $\sqcup S$, designated by $\vee$, which will have every state as a part.

Given an E-space $S = (S, P, \sqsubseteq)$, let $S' = S \upharpoonright P$ (the corresponding P-space) be the restriction $(P, \sqsubseteq \cap P^2)$ of S to P (where the first two components S and P within S are, in effect, identified). We also say in this case that S is an *E-extension* of S' and that S' is a *P-restriction* of S .

Lemma If $S = (S, P, \sqsubseteq)$ is an E-space then $S' = S \upharpoonright P$ is a P-space.

Proof We need to verify BC for S' . Suppose $Q \subseteq P$ has an upper bound t in P . Then it has a lub t' in S . Since $t' \sqsubseteq t$, $t' \in P$; and it is readily verified that t' is also a lub of Q in S' .

Completions

We wish to show how to extend a P-space to an E-space, in which fusions are always defined. To this end, we identify the states in the extended space with ideal-like objects. A subset I of states from a P-space $S = (S, \sqsubseteq)$ is said to be a (*virtual*) *ideal* if it satisfies the following two conditions:

- (i) *Upward Closure* any fusion of members of I belongs to I if it exists; and
- (ii) *Downward Closure* any part of a member of I is a member of I .

Note that (i) implies that $\wedge \in I$ since $\wedge$ is the fusion of the null set. We should also note that the fusion of all members of an ideal may not exist. If we form the restriction $\sqsubseteq'$ of $\sqsubseteq$ to the ideal I , then $(I, \sqsubseteq')$ will itself be a P-space. Thus the ideals correspond to subspaces of the given space.

Where T is a subset of S , let us use $T \uparrow$ for the *upward closure* of T , i.e. for the smallest superset of T to satisfy Upward Closure, and let us use $T \downarrow$ for the *downward closure* of T , i.e. for the smallest superset of T to satisfy Downward Closure. Thus (i) says $T \uparrow \subseteq T$ while (ii) says $T \downarrow \subseteq T$. Given a set of states T , we let T^+ be the smallest ideal to contain T (which is readily shown to exist).

We take $I[s]$ to be $\{t \in S : t \sqsubseteq s\}$. It is readily verified, given condition BC, that $I[s]$ is indeed an ideal. $I[s]$ is said to be the *principal ideal on s* ; and I is said to be a *principal ideal* if it is a principal ideal on some state and is otherwise said to be a *non-principal ideal*.

Intuitively, we think of ideals as fusions of the states that they contain (even though, strictly speaking, the fusion may not exist). An ideal is said to be *possible* or *consistent* when the fusion of all of its members exists and is otherwise said to be *impossible* or *inconsistent*. It is easily seen that an ideal is consistent iff it is principal.

Given a P-space $S = (S, \sqsubseteq)$, we let the (*mereological*) *completion* S^+ of S be $(S^+, P, \sqsubseteq^+)$, where S^+ is the set of ideals of S , P is the set of principal ideals of S , and $\sqsubseteq^+$ is defined on S^+ by:

$$s \sqsubseteq^+ t \text{ iff } s \subseteq t.$$

We prove two basic results on mereological completions.

Theorem 1 Given that $S = (S, \sqsubseteq)$ is a P-space, its completion $S^+ = (S^+, P, \sqsubseteq^+)$ is an E-space.

Proof It is evident that $\sqsubseteq^+$ is a partial ordering on S^+ . Also, the space $(S^+, \sqsubseteq^+)$ is complete. For take any subset T of ideals from S^+ . Let $t = (\cup T)^+$. Then $t \in S^+$ and is readily shown to be the lub of T . Finally, P , the set of principal ideals, satisfies Downward Closure. For suppose $I \sqsubseteq^+ [s]$, i.e. $I \subseteq [s]$. Then s is an upper bound of I and so $I = [\cup T]$.

We should note from the proof of this result that the fusion $I_1 \sqcup^+ I_2 \sqcup^+ \dots$ of ideals in S^+ is $(I_1 \cup I_2 \cup \dots)^+$, the closure of their union.

The second of our results says that completions embed the state structures from which they derive.

Theorem 2 (Embedding) Let $S = (S, \sqsubseteq)$ be a P-space and $S^+ = (S^+, P, \sqsubseteq^+)$ its mereological completion. Then the map I taking each state s of S to its principal ideal $I[s]$ is an isomorphism between S and the P-restriction $S^{+'} = (P, \sqsubseteq^{+'})$ of S^+ .

Proof By the definition of P in S^+ , I maps S onto P . I is one-one. For suppose $I[s] = I[t]$. Then, since $s \in I[s]$, $s \in I[t]$ and so $s \sqsubseteq t$; similarly, $t \sqsubseteq s$; and so $s = t$. Finally, we should show that for $s, t \in S$:

$$s \sqsubseteq t \text{ iff } I[s] \subseteq I[t].$$

If $s \sqsubseteq t$, then any member of $I[s]$, and hence part of s , will be a part of t , and hence a member of $I[t]$. Conversely, if $I[s] \subseteq I[t]$ then, since $s \in I[s]$, $s \in I[t]$ and so $s \sqsubseteq t$.

The mereological completion S^+ of S , as defined above, is not an extension of S , since each state s of S has been replaced by its principal ideal $I[s]$. To get an extension, we may replace each principal ideal $I[s]$ by its generating state s . We use a form of this alternative construction below.

Uniqueness

We show that the mereological completion is, in a certain sense, unique and generalize our results.

Say that a state within an E-space $S = (S, P, \sqsubseteq)$ is *P-based* if it is the fusion of consistent states. Clearly, given that a *P*-based state is the fusion of consistent states, it will be the fusion of all of its consistent parts. We say that the E-space $S = (S, P, \sqsubseteq)$ itself is *differentiated* if each of its states is *P*-based. Given a state s within an E-space $S = (S, P, \sqsubseteq)$, we let its *P-basis* $P_S(s)$ —or $P(s)$, when S is understood—be $\{t \in P: t \sqsubseteq s\}$. More generally, we may say that the subset Q of P is a *basis for S* if every state s of S is a fusion of members of Q . Thus an E-space $(S, P, \sqsubseteq)$ is differentiated if it has P as a basis.

Lemma 3 Let s and t be states within a differentiated space $S = (S, P, \sqsubseteq)$. Then:

- (i) Each *P*-basis $P(s)$ is a ideal
- (ii) $s \sqsubseteq t$ iff $P(s) \subseteq P(t)$
- (iii) $s = t$ iff $P(s) = P(t)$.

Proof (i) If u is a member of $P(s)$ then so is any part of u , given that P is downward closed. Suppose now that T is a subset of $P(s)$ and has a consistent fusion t . Then, since t is the lub of T , t is also a member of $P(s)$.

(ii) Clearly, $s \sqsubseteq t$ implies $P(s) \subseteq P(t)$. Now suppose $P(s) \subseteq P(t)$. Then $\sqcup P(s) \sqsubseteq \sqcup P(t)$. But $s = \sqcup P(s)$ and $t = \sqcup P(t)$; and so $s \sqsubseteq t$.

(iii) Similar to (ii) but with $=$ in place of $\sqsubseteq$.

Part (iii) of the lemma tells us that the states of a differentiated space can be individuated in terms of their consistent parts; two states will be the same when their consistent parts are the same.

Under certain conditions, we have a simple test for when an E-space is differentiated. Say that a state s in an extended space S is *thoroughly inconsistent* or *impossible* if it is inconsistent and yet contains no consistent non-null part. We have a *sheer* impossibility, so to speak, which cannot be attributed to any inconsistency among its parts. The following familiar condition may be imposed on the states within an E-space $S = (S, P, \sqsubseteq)$:

Strong Supplementation whenever s is a proper part of t there is a non-null part of t that is disjoint from s .

Theorem 4 Given Strong Supplementation, an E-space $S = (S, P, \sqsubseteq)$ is differentiated iff it contains no thoroughly impossible state.

Proof Suppose first that $S = (S, P, \sqsubseteq)$ contains a thoroughly impossible state s . Then the only consistent part of s is the null state and, for s to be the fusion of its parts it would itself have to be the null state and hence be consistent.

Suppose now that the state space $S = (S, P, \sqsubseteq)$ is not differentiated. Then some state s is not the fusion t of its consistent parts; and so t is a proper part of s . By Strong Supplementation, some state u is a part of s yet disjoint from t . But u is then thoroughly impossible for if u contained a consistent non-null part, that part would also be a part of t and so t and u would not be disjoint after all.

We say that f is a *standard* isomorphism between $S = (S, P, \sqsubseteq)$ and $S' = (S', P, \sqsubseteq')$ if it is an isomorphism between S and S' that is an identity on P (the identity of possible states is preserved):

Theorem 5 Suppose that $S = (S, P, \sqsubseteq)$ and $S' = (S', P, \sqsubseteq')$ are two differentiated extensions of the P-space $S = (P, \sqsubseteq)$ and that f is a one-one map from S onto S' . Then f is a standard isomorphism between S and S' iff

$$(*) f(s) = t \text{ iff } P_S(s) = P_{S'}(t) \text{ for each } s \in S \text{ and } t \in S'.$$

Moreover, if f is a standard isomorphism between S and S' , it is the only such isomorphism.

Proof Suppose first that f is a standard isomorphism between S and S' . The left to right direction of $(*)$ is then evident given that f is an identity on P . Suppose now that $P_S(s) = P_{S'}(t)$. Given that S is differentiated, $s = \sqcup P_S(s)$; and so, since f is an identity on P , $f(s) = \sqcup' P_S(s) = \sqcup P_{S'}(t)$. But, given that S' is differentiated, $t = \delta' P_{S'}(t)$; and so $f(s) = t$.

Suppose next that $(*)$ holds. We then have that:

$$s \sqsubseteq t \text{ iff } P_S(s) \subseteq P_S(t) \text{ by lemma 3(ii)}$$

$$\text{iff } P_{S'}(f(s)) \subseteq P_{S'}(f(t)) \text{ by } (*)$$

$$\text{iff } f(s) \sqsubseteq' f(t) \text{ again by lemma 3(ii).}$$

Finally, suppose that there was another standard isomorphism g between S and S' . Then for some $s \in S$ and distinct $t, u \in S'$, $f(s) = t$ and $g(s) = u$. But then $P_S(s) = P_{S'}(t)$ and $P_S(s) = P_{S'}(u)$; and so $P_{S'}(t) = P_{S'}(u)$, which is impossible given that S' is differentiated.

Given this theorem, there will be a canonical way of specifying any differentiated extension of a P-space $S = (S, \sqsubseteq)$. For let I be a set of non-principal ideals of S that is upward closed in the sense that, for any $J \sqsubseteq I$, $(\cup I)_+ \in I$. We then let $S + I$ be the space $(S^+, S, \sqsubseteq^+)$, where:

$$S^+ = S \cup I, \text{ and}$$

$$\sqsubseteq^+ = \sqsubseteq \cup \{(s, I): s \in S, I \in I \text{ and } s \in I\} \cup \{(I, J): I, J \in I \text{ and } I \subseteq J\}.$$

(This construction will only have its intended meaning if S itself contains no ideals of S ; and we shall assume this in what follows.)

Corollary 6 Given a P-space $S = (S, \sqsubseteq)$ and an upward closed set I of ideals of S , $S + I$ is a differentiated extension of S and any differentiated extension of S is isomorphic to the space $S + I$.

Proof It is readily verified that $S + I$ is an extension of S and its being differentiated follows from the fact that $P_{S+I}(I) = I$. Given a differentiated extension $S^+ = (S^+, S, \sqsubseteq^+)$, let $I = \{I: I = P_S(s) \text{ for some } s \in S^+ - S\}$. Using the above theorem, it is readily shown that $S + I$ is isomorphic to S^+ .

We say that f is a *standard* embedding of $S = (S, P, \sqsubseteq)$ into $S' = (S', P, \sqsubseteq')$ if it is an embedding which is an identity on P . Two differentiated extensions are of special interest:

Corollary 7 Given a P-space $S = (S, \sqsubseteq)$ there is (i) a minimal differentiated extension of S , i.e. one that is standardly embeddable in every other differentiated extension of S and there is (ii) a maximal differentiated extension of S , i.e. one in which every other differentiated extension is standardly embeddable.

Proof (i) Let I be the maximal ideal S . We may assume that I is non-principal since otherwise the minimal differentiated extension will be $(S, S, \sqsubseteq)$. Since the singleton set of ideals $\{I\}$ is upward closed, $S + \{I\}$ is a differentiated extension of S . Moreover, if $S^+ = (S^+, S, \sqsubseteq^+)$ is a

differentiated extension of S , it will contain a maximal element s for which $P_S^+(s) = S$; and it can then be shown that $S + \{I\}$ is standardly isomorphic to the restriction of S^+ to $S \cup \{s\}$.

Likewise, we may let I be the set of all non-principal ideals of S . Since I is upward closed, $S + I$ is a differentiated extension of S . Moreover, if $S^+ = (S^+, S, \sqsubseteq^+)$ is a differentiated extension of S , then it is readily shown to be standardly isomorphic to the restriction of $S + I$ to $S \cap \{I: I = P_S^+(s) \text{ for some } s \in S^+ - S\}$.

Factoring

We show how, under suitable conditions, an inconsistent state can be naturally divided into consistent parts.

Let $S = (S, \sqsubseteq)$ be a P-space, I an ideal of S , and s a state from I . Then s is said to be *maximal in I* if every state of I is either a part of s or incompatible with s . The maximal ideal of S is the set S itself; and so a state s will be maximal in the maximal ideal of S just in case any state whatever is either a part of s or incompatible with s . But this is just the definition of a world-state, given below. Thus we might think of maximal states within an ideal I as ‘mini-worlds’, relative to the space of states as defined by I .

We shall need the following condition on a P-space $S = (S, \sqsubseteq)$:

Ascent Every chain $s_1 \sqsubseteq s_2 \sqsubseteq \dots$ of states in S has an upper bound.

We use I^m for the set of maximal states in the ideal I . We have the following elementary results on maximal states:

Lemma 8 Let $S = (S, \sqsubseteq)$ be a P-space satisfying Ascent. Then:

- (i) Any state in an ideal is part of a maximal state in the ideal;
- (ii) Any ideal I is the downward closure of the set of its maximal states;
- (iii) Ideals containing the same maximal states are the same.

Proof (i) Proved in the usual way from Zorn’s Lemma (with the help of Ascent).

(ii) We need to show $I = I^m \downarrow$. If $s \in I^m \downarrow$, then s is part of some maximal element of I and so $s \in I$ by the downward closure of I . If $s \in I$, then s is part of a maximal element of I by (i) and so $s \in I^m \downarrow$.

(iii) Suppose that I and J have the same sets of maximal elements I^m and J^m . Then the downward closures $I^m \downarrow$ and $J^m \downarrow$ of these sets are the same and so, by (ii), I and J are the same.

It is important to note that (i) (and also the other clauses) may not hold if the given P-space $S = (S, P, \sqsubseteq)$ does not satisfy Ascent. For suppose S_ω simply consists of an ascending ω -chain $s_1 \sqsubseteq s_2 \sqsubseteq \dots$ of states. Then the ideal $S = \{s_1, s_2, \dots\}$ will have no maximal member.

Theorem 9 (Factorization) Let $S = (S, P, \sqsubseteq)$ be a differentiated E-space subject to Ascent. Then any state of S is the fusion of its maximal consistent parts.

Proof Without loss of generality, we may replace S with the corresponding space of ideals $S^+ = (S^+, P^+, \sqsubseteq^+)$. But from (ii) above, any state I of S^+ will be the fusion of $\{t\}$ for t maximal in I and hence I will be the fusion of its maximal consistent parts.

Modal Completion Let $S = (S, P, \sqsubseteq)$ and $S' = (S', P', \sqsubseteq')$ be two E-spaces. We define their *product* $S \times S'$ to be the structure $(S \times S', P \times P', \sqsubseteq^*)$, where $\sqsubseteq^*$ is defined pointwise as $\{(s, s'), (t, t')\}: s \sqsubseteq t \text{ and } s' \sqsubseteq' t'\}$. It is readily shown:

Lemma 10 Given that S and S' are E-spaces, then so is their product $S \times S'$.

A state from a state space is said to be *necessary* if it is compatible with every consistent state. Given an E-space $S = (S, P, \sqsubseteq)$, we use $S_\square$ for the set of its necessary states and take the *modal complement* S_m of S to be the structure $(S_\square, \{\wedge\}, \sqsubseteq_\square)$, where $\sqsubseteq_\square$ is the restriction of $\sqsubseteq$ to $S_\square$.

Intuitively, we think of each member s of $S_{\square}$ in S as representing its opposite s , which is why $\wedge$ is taken to be the only possible state in $S_{\square}$.¹⁸

Every part of a necessary state of an E-space S will also be a necessary state. But the fusion of necessary states may not be a necessary state and it may not even be consistent, as is illustrated by the completion of the space S_{ω} above, since all of its consistent states $s_1, s_2, \dots$ are necessary and yet lack a consistent fusion.

Say that an E-space S is $\square$ -complete if $\sqcup S_{\square} \in S_{\square}$. Then we readily show:

Lemma 11 If the E-space S is $\square$ -complete, then its modal complement S_m is an E-space.

Under a certain natural condition, the space S will be $\square$ -complete. Say that a state s of S is a *world-state* if it is consistent and if any consistent state is either a part of s or incompatible with s ; and say that the space S is a *W-space* if every consistent state of S is part of a world-state.

Lemma 12 Every W-space is $\square$ -complete.

Proof First establish:

(*) a state is necessary iff it is a part of every world-state.

It follows from (*) that every member of $S_{\square}$ is a part of every world-state and therefore that $\sqcup S_{\square}$ is a part of every world-state and hence necessary, by (*) once again.

We now give the construction. Given a $\square$ -complete E-space $S = (S, P, \sqsubseteq)$, we let its *modal completion* $S^m = (S^m, P^m, \sqsubseteq^m)$ be the product $S \times S^m$ of S and its modal complement S^m . Thus each state of S^m is of the form (s, t) , with $s \in S$ and $t \in S^m$. Intuitively, we think of (s, t) as representing the fusion of s and t . Since each of S and S^m is an E-space, the modal completion S^m will also be an E-space.

The modal completion S^m may not be differentiated even if S is differentiated since distinct s and t , with $s, t \in S_{\square}$ will have the same consistent states as parts. However, each modal completion will have $\{(s, \wedge): s \in S\} \cup \{(\wedge, s): s \in S_{\square}\}$ as a basis and hence will have $\{(s, \wedge): s \in P\} \cup \{(\wedge, s): s \in S_{\square}\}$ as a basis as long as the original space S is differentiated.

Semantics

For simplicity, we deal with the case of truth-functional logic, although the semantics can also be extended to quantificational logic.

Formulas of the language L of truth-functional logic are constructed from the sentence-letters $p_1, p_2, p_3, \dots$ and the connectives $\neg, \wedge$ and $\vee$ in the usual way.

A pair of sets of states (V, F) within a P-space $S = (S, P, \sqsubseteq)$ is said to be a *putative verification-falsification condition* (or *putative VF-condition*, for short) Intuitively, we think of V and F as possible verification- and falsification conditions for a given statement. Where $G = (V, F)$ is a putative VF-condition, we use $[G]^+$ for V and $[G]^-$ for F .

We say that the putative VF-condition (V, F) is:

exclusive if no state in V is compatible with a state in F ;

exhaustive if any state is compatible with a state in V or with a state in F .

A putative VF-condition is then said to be an (*genuine*) *VF-condition* if it is both exclusive and exhaustive.

A *P-model* M for the language L is an ordered triple $(S, \sqsubseteq, [\cdot])$, where $(S, \sqsubseteq)$ is a P-space and $[\cdot]$ (valuation) is a function taking each sentence letter into a VF-condition. Given a state model $M = (S, \sqsubseteq, [\cdot])$, we may define what it is for a formula A to be (*exactly*) *verified* by a given state s ($s \models A$) or to be (*exactly*) *falsified* by the state s ($s \models A$):

¹⁸ For certain purposes, we might wish for $\wedge$ to have a complement, in which case we would have to add a new state to serve as the null state for S_m .

- (i)⁺ $s \Vdash p$ if $s \in [p]^+$;
- (i)⁻ $s \nVdash p$ if $s \in [p]^-$;
- (ii)⁺ $s \Vdash \neg B$ if $s \nVdash B$;
- (ii)⁻ $s \nVdash \neg B$ if $s \Vdash B$;
- (iii)⁺ $s \Vdash B \wedge C$ if for some t and u , $t \Vdash B$, $u \Vdash C$ and $s = t \sqcup u$;
- (iii)⁻ $s \nVdash B \wedge C$ if $s \nVdash B$ or $s \nVdash C$;
- (iv)⁺ $s \Vdash B \vee C$ if $s \Vdash B$ or $s \Vdash C$;
- (iv)⁻ $s \nVdash B \vee C$ if for some t and u , $t \nVdash B$, $u \nVdash C$ and $s = t \sqcup u$.

Note that clause (iii)⁺ for conjunction requires that the verifier $t \sqcup u$ for $B \wedge C$ should exist; and similarly for clause (iv)⁻.

The current semantics is readily extended to an E-space $(S, P, \sqsubseteq)$. The definition of a putative VF-condition (V, F) is the same as before but now the states used to define exclusivity and exhaustivity should be taken to be consistent. Thus a putative VF-condition (V, F) will be:

- exclusive* if no consistent state in V is compatible with a consistent state in F ;
- exhaustive* if any consistent state is compatible with a state in V or with a state in F .

An *E-model* M for the language L is an ordered quadruple $(S, P, \sqsubseteq, [\cdot])$, where $(S, P, \sqsubseteq)$ is an E-space and $[\cdot]$ (valuation) is a function taking each sentence letter into a VF-condition. The semantic clauses (i)–(iv) are then exactly the same as before. But note now that the state $s = t \sqcup u$ in clauses (iii)⁺ and (iv)⁻ is guaranteed to exist by the completeness of the space $(S, P, \sqsubseteq)$.

Say that the E-model $M = (S, P, \sqsubseteq, [\cdot])$ is a *conservative extension* of the P-model $M' = (S', \sqsubseteq', [\cdot]')$ if $S' = P$, $\sqsubseteq' = \sqsubseteq \upharpoonright P$, and $[\cdot]' = \{(V \cap P, F \cap P) : (V, F) \in [\cdot]\}$. The following result is established by a straightforward induction:

Theorem 13 Suppose that the E-model $M = (S, P, \sqsubseteq, [\cdot])$ is a conservative extension of the P-model $M' = (S', \sqsubseteq', [\cdot]')$. Then for any formula A and any state $s \in P$,

- (i)⁺ $s \Vdash A$ in M iff $s \Vdash A$ in M' , and
- (i)⁻ $s \nVdash A$ in M iff $s \nVdash A$ in M' .

This result shows that the E-semantics will differ from the P-semantics simply in allowing a statement to have certain inconsistent verifiers or falsifiers that it did not have before. An analogous result holds for other versions of truthmaker semantics. What is essential is this: treating a connective, say $\wedge$, as a function f from two sets X and Y of VF-conditions from a P-model into a set $f(X, Y)$ of VF conditions, the extension of the function to an E-model should be ‘conservative’ over the consistent states, i.e. if the pair X and X' and the pair Y and Y' agree with respect to the consistent states, then so should the pair $f(X, Y)$ and $f(X', Y')$.

References

- Berto, F. (2009) Impossible worlds., *Stanford Encyclopedia of Philosophy*.
- Brandom R. and Rescher N. (1980) *The Logic of Inconsistency*. Oxford: Blackwell.
- Edgington D. (1985) The Paradox of Knowability., *Mind* 94: 557–68.
- Edgington D. (1995) On Conditionals., *Mind* 104: 235–329.
- Edgington D. (2010) Possible knowledge of unknown truths. *Synthese* 173(1): 41–5.
- Fine K. (2012a) A difficulty for the possible worlds analysis of counterfactuals., *Synthese* 189(1): 29–57.
- Fine K. (2012b) Counterfactuals without possible worlds., *Journal of Philosophy*, 109(3): 22–46.

- Fine K. (2014) Truth-maker semantics for intuitionistic logic, *Journal of Philosophical Logic* 43(2–3): 549–77.
- Fine, K. (2016) Angellic content, *Journal of Philosophical Logic* 45(2): 199–226.
- Fine (2017a) A Theory of Truthmaker Content I, *Journal of Philosophical Logic* 46 (6): 625–674.
- Fine, K. (2017b) A theory of truthmaker content II: Subject-matter, common content, remainder ground, *Journal of Philosophical Logic* 46(6): 675–702.
- Fine, K. (2017c) Truthmaker semantics. Chapter 22 in B. Hale, C. Wright, and A. Miller (eds), *A Companion to the Philosophy of Language* (2nd edn), Chichester: Wiley & Sons Ltd, pp. 556–77.
- Fine, K. (2018) Yablo on subject matter, *Philosophical Studies*, 1–43.
- Gemes K. (1994) A New Theory of Content., *Journal of Philosophical Logic* 23, 596–620.
- Gemes K. (1997) A New Theory of Partial Content II: Model Theory and Some Alternatives., *Journal of Philosophical Logic* 26, 449–476.
- Rescher, N. & Brandom, R. B. (1980) *The Logic of Inconsistency: a Study in Non-standard Possible Worlds Semantics and Ontology.*, Oxford: Basil Blackwell.
- Restall, G. (1997) Ways Things Can't Be., *Notre Dame Journal of Formal Logic*, 38: 583–96.
- Van Fraassen B. (1969) Facts and Tautological Entailments., *Journal of Philosophy* 66, 477–87.
- Yablo S. (2014) *Aboutness*, Princeton University Press.

The Epistemic Use of ‘Ought’

John Hawthorne

A good deal of Dorothy Edgington’s work has involved fruitful applications of the probability calculus to philosophical subject matters—notably, conditionals and vagueness. This chapter forms part of a project of exploring the relevance of probability to various epistemic phenomena, including knowledge and epistemic modality. My focus here is on certain epistemic uses of ‘ought’ and ‘should’. I argue against flat-footed ways of grounding those concepts in the ideology of probability, although make room for certain other, less reductive, structural relationships between the two.

It has often been noticed that there is an apparently epistemic use of the modal auxiliary ‘ought’ (and also ‘should’) whereby ‘Ought P’ conveys some positive epistemic status for the proposition that P.¹ An utterance of

- (1) Jim ought to be in London by now

may be devoid of ethical content (Jim may be an escaping criminal whose presence in London is a morally bad thing). Nor need it be a way of expressing what is conducive to Jim’s desires or be in some other way prudential.² Instead, the utterance may have the effect of epistemically warming the audience to the hypothesis that Jim is in London. What is the meaning of ‘ought’ (and ‘should’) in contexts like this?

10.1 Proposal One

A quite natural suggestion—one that I often encounter in philosophical conversation and which straightforwardly accounts for the positive epistemic force of ‘ought’ in the relevant contexts—is that ‘ought’, so used, means the same as ‘probably’ (or ‘very probably’ or ‘it is highly probable that’) in its epistemic use. Ralph Wedgwood is representative here:

...there is the epistemic ‘ought’, as in ‘Tonight’s performance ought to be a lot of fun’, which seems to mean, roughly, just that it is highly *probable* that tonight’s performance will be a lot of fun...

¹ I shall assume without argument that ‘ought’ functions semantically as a propositional operator.

² As in ‘If Jim wants to see the football match, he ought to get to London by 2pm.’

... an epistemic 'ought'-statement, of the form 'In relation to evidence *E*, it ought to be that *p*' is true if and only if *p* is sufficiently probable given evidence *E*.³

Something worth noticing immediately is that this idea does not sit happily with an intuitively compelling principle for the logic of 'ought' in all its myriad uses, namely:

Agglomeration: (Ought *P* and Ought *Q*) $\supset$ Ought (*P* and *Q*)

The validity of the agglomeration schema falls out of an increasingly popular account of the barebones logical structure of 'ought'—associated above all with the work of Angelika Kratzer—according to which any use of 'ought' is associated with a domain of worlds (the 'modal base') and a ranking on those worlds (derived from an 'ordering source', which is a function from sets of worlds to rankings of the worlds in that set). Relative to a domain and ranking, 'Ought *P*' is true just in case there is some world such that any world ranked equal to or ahead of it is a *P* world. (In the case of a finite domain this is equivalent to the requirement that all of the top ranked worlds are *P* worlds.⁴) Kratzer's approach to modals has been incredibly fruitful, offering a satisfying vision of the structural commonalities of 'must', 'ought', and so on in their myriad uses. As someone who finds Agglomeration intuitively compelling and who is, moreover, also impressed by the semantics sketched above, the probability gloss strikes me as problematic. After all, on no reasonable construal of 'Probably' does the schema:

(Probably *P* and Probably *Q*) $\supset$ Probably (*P* and *Q*)

come out as valid. (The same point applies to various meanings of 'very probable', 'sufficiently probably', and so on).⁵

Agglomeration is not the only challenge to the probability gloss. There are apparently legitimate uses of this kind of 'ought' where it is clear that the evidence against *P* is overwhelming in the setting in which the claim is made. Thus there are uses of

(2) Jim ought to be in London by now but isn't

that appear to involve much the same use of 'ought', uttered in a setting where the speaker knows that the proposition that the train has arrived is false.⁶ (This does not itself constitute a decisive objection to a probability based approach. One might

³ Wedgwood (2006).

⁴ See, for example, Kratzer (2002) and 'Modality' in von Stechow and Wunderlich (2008). Note that the semantics stated in the text does not vindicate a version of agglomeration for countable sets of propositions: if the modal base is infinite and every world has some world ranked above it by the operative ordering source, then one can contrive a countable set *P*₁, *P*₂ ..., where 'Ought *P*₁', 'Ought *P*₂' ... comes out true but where Ought *C* (where *C* is the infinite conjunction of *P*₁ ... *P*_n) is not true. I do not wish to take a stand on 'infinite agglomeration' here.

⁵ I mention one defensive move available to the proponent of proposal one. Call an inference form 'OKish' if it never takes one from definitely true premises to a definitely false conclusion. Suppose 'Ought' means 'sufficiently probably' and that 'sufficiently probably' is inevitably rather vague at any context. Then it may be that the inference form: 'Ought *P*. Ought *Q*. Therefore Ought (*P* and *Q*)' is OKish. This doesn't strike me as a satisfactory rejoinder but perhaps it helps to soften the blow for one who gives up finite agglomeration.

⁶ Thanks to Eric Swanson here.

think that 'ought' on this use means something roughly like 'Probable relative to the contextually relevant evidence', and insist that in the context of utterance of (2), the proposition that Jim is not in London is not part of the relevant evidence. But this kind of consideration certainly puts pressure on a probability based approach.)

There are further concerns. Suppose I am about to throw a die. The claim

- (3) A number between one and four ought to come up.

does not seem felicitous. But a gloss on 'ought' as 'probably' would predict felicity. One might try to fix this by construing 'ought' as 'very probably', which is in turn tied to a threshold set by context. But notice that the claim 'He ought to lose the lottery' does not sound like much of an improvement even if the odds of winning are fantastically low. Here a contrast deserves emphasis. The claim

- (4) If I throw a die fifty times, a number between one and four ought to come up sooner or later

sounds perfectly felicitous. But the claim 'He ought to lose the lottery' sounds far less felicitous, no matter how vast the number of tickets. A probability-based semantics cannot explain this.⁷

10.2 Proposal Two

A second proposal that I have encountered connects the 'ought' of interest with 'ought to believe'. The idea here is that 'It ought to be that P' is equivalent to 'Group g ought to believe P', where the group (which may consist of a single individual, perhaps the speaker himself) is set by context. (Of course there then remains the task of analysing 'ought to believe'.) This is not so obviously incompatible with Agglomeration, since it is at least arguable that insofar as one ought to believe P and one ought to believe Q then one ought to believe P and Q. And this idea can potentially explain why (3) above is infelicitous, since it is arguable that one ought not flat out believe (3) in the situation envisaged. (This is especially clear if one thinks of knowledge as the norm of flat out belief.) But if this is used as the explanation of the awkwardness of (3), it is hard to understand why one can happily use 'It ought to be that P' in all manner of contexts where it is perfectly clear that one is in no position to know P and where one is reluctant to undertake commitment to flat out belief. Consider a natural setting in which one says:

- (5) He certainly ought to have arrived by now but we had better check.

This is not, intuitively, a setting in which one thinks that one ought to flat-out believe that he has arrived prior to checking. Now of course if one glossed the relevant 'ought to believe' claim as something like 'One ought to be highly confident' or 'One ought to have high subjective probability', then that would sit more easily with a willingness

⁷ I note in passing that while it is very easy to get an epistemic reading of 'Jim is probably not in London', most informants that I asked found it very difficult indeed to get an epistemic reading of 'He ought not to be in London' and 'He should not/shouldn't be in London.' I don't have a good explanation of why epistemic uses of these terms do not sit happily with 'not'.

to assert (5). But, as can be seen from the discussion of Proposal One, such a gloss would make Agglomeration hard to sustain and would not explain the unassertability of claims like (3). Such a reading would also predict gross infelicity for:

- (6) He certainly ought to have arrived by now but there are so many problems with Virgin trains these days that it is hard to be very confident that he has.

But this speech does not sound particularly awkward. Similarly, whether one glosses belief as a committal state that aims at knowledge or as mere high confidence, the 'ought to believe' idea does not sit well with (2)—in those settings it is perfectly clear that one ought to believe not-P even though one says 'It ought to be that P'. (I suppose one might try to accommodate this by taking the relevant group to be a hypothetical group which lacks some of the speaker's evidence, but that seems like a rather desperate move.)

Note finally that the 'ought to believe' gloss does not extend happily to certain embedded uses of 'It ought to be that P'. Suppose I say 'I have no idea which train he is on, but if he is on the fast train he ought to be there by now.' If we replace the consequent by an 'ought to believe claim' we get: 'I have no idea which train he is on, but if he is on the fast train I/we ought to believe that he is there by now.' But this latter sentence sounds much less felicitous: if the person is on the fast train the speaker nevertheless has no good evidence that this is so.

10.3 Proposal Three

Unlike the previous two proposals, a suggestion raised in passing (though not endorsed) by Kai Von Fintel does secure Agglomeration. The idea to capture the content of 'ought' in these contexts by using relative plausibility as an ordering source on worlds.⁸ 'Ought P' will then, as is usual in a Kratzer-style semantics, come out true relative to an ordering source just in case there is a P world such that no world ranked equal or higher than it is a not-P world. An approach that ranks worlds by probability or plausibility and which then applies the standard Kratzer rule works very differently to a gloss on 'Ought' as simply meaning 'Plausible' or 'Probable'. For if one ranks by relative plausibility/probability and then applies the rule articulated above, Agglomeration will be satisfied (since the logic of the rule guarantees Agglomeration relative to any ordering source).

To handle data sentences of the form 'He ought to be here by now but he isn't' this approach will have to make certain moves analogous to those of the probability approach. Thus it might be suggested that we sometimes rank worlds not in terms of what is plausible relative to our total body of evidence but on what is plausible relative to some salient less-than-total range of evidence.

The key problem with this suggestion, however, is that it overlooks the fact that quite often the most plausible world will not be particularly plausible at all.⁹ If one

⁸ See 'Modality and Language' in Borchert (2006), p. 4 where 'plausibility or stereotypicality' are mentioned as candidate ordering sources for epistemic modality.

⁹ I realize this may have been a bit of a throwaway remark by Von Fintel – my concern here is not to criticize him but to make sure readers are clear about why this approach does not work.

ranks worlds for plausibility and applies the suggested semantics, it can easily turn out that 'Ought P' is true even though it is far more plausible that not-P than P.

Suppose for example we are wondering whether it will snow, rain, be cloudy (but with no rain), or be sunny. For simplicity let us suppose there is just a single world corresponding to each hypothesis. Suppose it is 35 per cent likely that it will be sunny and that each of the other hypotheses is individually less than 35 per cent. Then the proposed semantics makes 'It ought to be sunny' come out true. But this does not seem like a good result. That sunniness has some incremental advantage over the other hypotheses hardly suffices for the truth of the relevant ought claim.

10.4 Proposal Four

Another proposal, endorsed by Von Fintel and Iatridou is far more compelling.¹⁰ It is that epistemic 'ought' uses relative normality as the ordering source. Let me highlight a few compelling features of this view.

First, the view has no problem with Agglomeration: if one ranks by relative normality and applies the Kratzer semantics sketched earlier, then Agglomeration will fall out.

Second, the view does very well with lottery data since it predicts that if, at a context, the large majority of possible scenarios are P entailing but that there are many non-P scenarios that seem just as normal, then 'Ought P' will not seem plausible. This is borne out by the lottery case: since my losing is not less normal than any other particular person with a ticket losing, the situation where I win is intuitively just as normal as any given one where I don't win. But then the scenarios that tie for maximal normality are not permeated by my not winning. On the other hand, normality considerations will make for the truth of 'Ought P' in certain settings where it is by no means more probable that P than it is that I will lose some lottery. Thus, returning to an earlier example, it is intuitively natural to count as abnormal worlds in which I throw a die fifty times but where a particular number never comes up. So even if I had a ticket in a vast lottery where the odds of winning are even less than the odds of a 1 failing to come up in fifty dice rolls, the claim 'A one ought to come up some time in the fifty rolls' sounds markedly better than 'I ought to lose the lottery.'

Third, the view offers a nice account of the commonality between (1) and (2). In both settings a normality ordering source is in play. From the perspective of an ordering source plus modal base semantics there may still be an important contrast, in that while the speaker's evidence plausibly provides the modal base for (1), the modal base of (2) includes worlds that are incompatible with the speaker's evidence. What this suggests is that the semantic relevance of evidence to epistemic 'oughts' has merely to do with the modal base—it does not infiltrate the way of ranking worlds. On this vision then, what gives certain uses of 'ought' a distinctively epistemic feel is not the kind of ordering source—normality ordering sources don't always have

¹⁰ See von Fintel and Iatridou (2008).

that feel—but instead the combination of a normality ordering source with an evidence-based way of setting the modal base.

Supposing that the ordering source for worlds is normality rather than anything directly epistemic, why do these ought claims often provide us evidence for the proposition within the scope of the 'ought'? There is nothing especially puzzling here. In general, if one learns that normal situations are P situations that provides evidence for P. Of course that evidence can be offset or even swamped by countervailing evidence, as in (6) and (2). (Note here that risks of abnormality will add up so that ought claims reached by repeated agglomeration, while true, will provide at best very weak evidence for the relevant conjunction.)

Fourth, the view correctly predicts two sources of context dependence for broadly epistemic ought's, one having to do with variation in the modal base from speaker to speaker, one having to do with shifty standards of normality. Suppose a train has been diverted from destination A to destination B. A speaker that does not know that can intuitively speak correctly by saying 'She ought to have arrived at A by now.' Another speaker who is apprised of the rerouting can intuitively say, 'She ought to have arrived at B by now.' This is context-dependence that is rooted in a varying modal base: the most normal worlds among a set that includes worlds where the train goes on its normal course will be one's where the person has arrived at A, but the most normal among a set populated by rerouting scenarios are ones where the person has arrived at B. Consider meanwhile a baker that two speakers know to be a perfectionist and, more often than not, throws a partly made cake in the bin prior to completion owing to salient imperfections. One speaker says 'You should find that there is a partly made cake in the bin.' Another says 'The cake ought to be made by now but I wouldn't be surprised if the baker has had a tantrum and thrown it away.' Here we have shifting standards of normality—one speaker focuses on the normal course of baking a cake, another one the normal practice of that baker. Note that one distinctive kind of context-dependence stems from the thought that it is normal for something abnormal to happen if there are enough instances of a process. Suppose a village has three thousand washing machines and it is washing day. Repeated applications of agglomeration yield, at a context, the true sounding 'There ought to be a pile of clean washing in each house by the end of the evening.' But in another context, one where one deploys standards of normality geared to the normal distribution of faults and breakdowns, it is natural to say 'There ought to be at least one house with a broken washing machine and where there won't be any clean clothes today.'¹¹

10.5 Proposal Five

Proposal four appears to break any systematic tie between epistemic uses of 'ought' and 'probably'. Given the semantics sketched there is no expectation that, at a context, 'Probably P' will entail 'Ought P', nor any expectation that 'Ought P' will

¹¹ Relativists will no doubt try to point to some disagreement data that allegedly favours relativist rather than contextualist semantics here. I shall not be pursuing those issues here.

entail 'Probably P'. But if so called epistemic uses of 'ought' have no systematic connection with 'Probably' this can make for some puzzling results. Suppose that at a context 'Jim ought to be in London' is associated with a modal base fixed by the evidence and a normality ordering source that ranks certain Jim in London worlds higher than other worlds. Suppose that for each epistemically possible time of arrival in London (suppose for example the speaker knows that if Jim did arrive it was between 5pm and 7pm) there are some worlds among the top ranked worlds where the speaker arrives at exactly that time—none is considered especially abnormal. Suppose the speaker then learns that either Jim arrived at exactly 6.03 in London or else broke down. And suppose it is very unlikely—though still possible—on the new set of evidence that Jim arrived at exactly 6.03, but that it is far more likely on new evidence that Jim is not in London than that he arrived at 6.03. Assuming that the kind of normality ranking in play remains the same, then it seems that the 6.03 world will be ranked higher than the breakdown worlds and so 'He ought to be in London' will be true in the updated context. But in a setting where one uses one's evidence to fix the modal base there is something a bit troubling about the claim 'He ought to be in London but it is overwhelmingly likely that he isn't.' Consider similarly the claim 'A six ought to come up sooner or later', said of a person about to throw a die 50 times. Suppose a normality ordering source is in play that ranks mixed distributions as more normal than distributions with uniform outcomes and so, in particular, ranks various sequence with 49 non-sixes followed by a six as more normal than 50 sixes in a row. Suppose the person is then told 'Either the first 49 or the first 50 throws were all non-sixes.' Suppose the same kind of normality ranking is still applied to the new modal base generated by the new evidence. Then the speech 'Well, the last throw ought to be a six' comes out true. But such a speech does not sound felicitous in that sort of context and a natural diagnosis of this infelicity is that we expect such uses of 'ought' to entail 'probably'.

The problem extends to conditionals. In a setting where the modal base is S, it is plausible that a speech of the form 'If P, Ought Q' has the effect of restricting the modal base to the subset of S compatible with P. (Thus we can say 'If Jim took the express train he ought to be in London by now' in a setting where it is more normal for Jim to economize and take the slow train. Here we need a restricted modal base for the purpose of evaluating the consequent.¹²) Suppose now that the ordering source ranks die throwing worlds along the lines described above. Then, uncomfortably, it seems in the original setting, the conditional 'If he doesn't throw a six the first 49 times, he ought to throw a six the 50th time' will come out true. But far from being true, such a speech seems like a paradigm instance of the gambler's fallacy.

Problems of this sort are a special case of a more general kind of problem, namely that of finding some comfortable marriage between probabilistic concepts thrown up by science and decision theory and epistemological concepts whose basic conceptual underpinnings are not probabilistic at all (here I have in mind the concept of knowledge, as well as the epistemic uses of modals such as 'might', 'must', and 'ought'). The various families of concepts are arguably not crafted out of the same

¹² For an extended treatment of these topics see Dorr and Hawthorne (2013).

conceptual cloth and it is not at all easy to move back and forth between them in a satisfying way. (Puzzles concerning the relationship of objective chance to knowledge are one instance of this.¹³) Perhaps no proposal is pain free. Yet can anything be done to craft a use of 'ought' that—as against the spirit of proposal four—crafts a close tie between 'Ought P' and 'Probably P' without sacrificing the underlying logic of 'Ought' (as proposal one unfortunately does)? Here is one idea that will help in this regard (by securing an 'ought' to 'probably' inference).¹⁴ Start with a normality ordering, O, on all the worlds.¹⁵ From this we can derive an ordering source S—namely a function from sets of worlds to orderings which simply ranks the worlds in the set according to O. But we can also derive what I shall call a 'flattened' normality ordering source based on O. For any set S₁, the flattened normality ordering source ranks as equal top the smallest set of worlds w₁ ... w_n in S₁ that satisfies the following conditions: (i) The worlds in question form a set S₂ that is more likely than not conditional on S₁ and (ii) every world in S₁ that is not in S₂ is ranked by O to be lower than any world in S₂.

For example, consider the three membered set of worlds w₁ w₂ and w₃ where O ranks w₁ ahead of w₂ which is in turn ranked ahead of w₃. Suppose each world is equiprobable conditional on the information that one of those three worlds obtains. The flattened normality source for that set will rank w₁ and w₂ equal top (it can't rank just w₁ top since that would violate condition (i)) and will rank w₁ and w₂ ahead of w₃. Consider meanwhile the subset of that three membered set consisting of just w₂ and w₃. O will rank w₂ ahead of w₃ in that set. By contrast the flattened normality source will not rank w₂ ahead of w₃ in that set (since doing so would again violate condition (i)). It is worth taking note of an important structural contrast between O and the flattened source based on O: If a world x is ranked higher than a world y by O in a given set to which they both belong, then x will be ranked higher than y by O in any superset and any subset to which they both belong. But this is not true of the flattened normality source. In the latter case, examples can be contrived where x gets ranked higher than y relative to some set z but only gets equal ranking relative to certain supersets and certain subsets. (For an example concerning subsets, see above. For an example concerning supersets consider first the trio above—where the flattened normality source ranks each of w₁ and w₂ as above w₃—and then a five membered superset consisting of w₁, w₂, and w₃ and two more worlds, each of which is ranked by O as below w₂ but on a par with w₃. Here the flattened normality source must rank w₃ equal top—along with the other four worlds—on pain of violating condition (i).)¹⁶

¹³ See Hawthorne and Lasonen-Aarnio (2009), together with Williamson's reply in that volume.

¹⁴ I shall not be trying to secure a 'probably' to 'ought' inference since we have seen that this is highly questionable. (See (3) for example.)

¹⁵ I am obviously well aware that in practice the ordering source deployed at a context will be beset by vagueness. I am idealizing away from issues of vagueness in this discussion. I am also in what follows ignoring issues of hyperintensionality by taking propositions to be sets of worlds and hence assuming that conditional probabilities can be coded by relative probability of one set of worlds on another. (This idealization is of course very standard in contemporary decision theory.)

¹⁶ Of course the flattened normality source will never reverse rankings so that some world x gets ranked above y relative to one set but y gets ranked below x relative to some other set. For a discussion of deontic

Suppose that a flattened normality source were available for certain uses of 'ought'. What effects would this induce? Notice first that at contexts where the modal base is set by the speaker's evidence and a flattened source is in play, the truth of 'Ought P' will guarantee the truth of 'Probably P'. After all, if P is true at all the top ranked worlds in the modal base it is at least as likely as the set consisting of the top ranked worlds, and the flattened normality source is crafted so that the latter set is more likely than not relative to the modal base. And insofar as a flattened source is in play, the puzzles raised in connection to proposal four will not arise. Consider the fifty die roll case. Suppose the normality ordering is as described. Then the flattened normality ordering for a set containing all the possible die rolls will exclude worlds in which no sixes are thrown from the top ranked set. But consider now a set S containing worlds consisting of all non-sixes and worlds where all but the last is a six and which are otherwise equally normal, and where the set consisting of the six worlds is not likely on S. The normality ordering source will rank the non-six worlds as below the six worlds relative to this set. But the flattened normality source will not rank the non-six worlds as lower, on pain of violating condition (i). Thus, if the information comes in that either 49 or 50 non-sixes have been thrown, and the modal base is shrunk to such a set, then the flattened normality source would not license 'The last one ought to be a six.' The infelicity of the conditional 'If he doesn't throw a six the first 49 times, he ought to throw a six the 50th time' can similarly be explained if a flattened ordering source is in play.

My discussion has involved various idealizations and simplifications. Nevertheless, it seems to me that flattened ordering sources are a useful tool for developing a broadly Kratzer-style treatment of at least some uses of 'ought'. Notice that for these uses, epistemic factors have a dual role in fixing the context: Evidence will on the one hand fix the modal base, and considerations of epistemic probability will shape the ordering source. I do not wish to pretend that this will effect a perfectly happy marriage between 'ought' and the concept of epistemic probability. But it is perhaps the best alliance that can be achieved given the broad logical structure of 'ought'. I by no means want to suggest either that all normality driven uses of 'ought' use flattened normality orderings. But the suggestion that some of them do—particularly ones with an especially epistemic feel—is worthy of serious consideration.

References

- Borchert, D. (ed.) (2006) *Encyclopedia of Philosophy*. Basingstoke: Macmillan.
- Dorr, C. and Hawthorne, J. (2013) Embedding epistemic modals, *Mind* 122: 867–914.
- Fintel, K. von and Iatridou, S. (2008) How to say 'ought' in foreign: the composition of weak necessity modals. In J. Guéron and J. Lecarne (eds), *Time and Modality*, Netherlands: Springer, pp. 115–41.
- Hawthorne, J. and Lasonen-Aarnio, M. (2009) Knowledge and objective chance. In P. Greenough and D. Pritchard (eds), *Williamson on Knowledge*, Oxford: Oxford University Press, pp. 92–108.

ordering sources with that effect see Kolodny and MacFarlane's discussion of 'serious information dependence' in section 4.3 of 'If' and 'Oughts' (Kolodny and MacFarlane, 2010).

- Kolodny, N. and MacFarlane, J. (2010) 'If' and 'oughts', *Journal of Philosophy* 107: 115–43.
- Kratzer, A. (2002) The notional category of modality. In P. Portner and B. Partee (eds), *Formal Semantics: The Essential Readings*, Oxford: Blackwell, pp. 289–323.
- von Stechow, A. and Wunderlich, D. (eds) (2008) *Semantics: An International Handbook of Contemporary Research*, Berlin: Walter de Gruyter, pp. 639–50.
- Wedgwood, R. (2006) The meaning of 'ought'. In R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*, vol. 1, Oxford: Oxford University Press, pp. 127–60.

11

Undercutting Defeat and Edgington's Burglar

Scott Sturgeon

11.1 Game Plan

This chapter does four things.* First it lays out an orthodox position on reasons and defeaters. Then it argues that the position just laid out is mistaken about 'undercutting' defeaters. Then the chapter explains an unpublished thought experiment by Dorothy Edgington. And then it uses that thought experiment to motivate a new approach to undercutting defeaters.

11.2 Reasons and Defeaters

Defeasible reasons are normally thought of as mental states of some kind. In the verbal tradition, at least, reputable philosophers sometimes react to this fact as if the whole idea of a defeasible reason is based on some kind of conceptual confusion or category mistake. Their idea, basically, is that the English word 'reason' already has a meaning which rules out mental states as part of its extension. For this reason they see the idea of mental states as reasons as itself utter confusion.

My view is that the meaning of the English word 'reason' is irrelevant to the debate about defeasible reasons; for the claim that defeasible reasons are mental states—to be made here and found in the literature—should be thought of as a matter of *legislation*: if you like, the phrase 'defeasible reason' should be understood as a technical one, something which picks out by fiat, if it picks out anything, whatever plays the role pinned down by it in theory; and the same is true, *mutatis mutandis*, for the word 'reason' in what follows.

* Material presented here is based on talks given at the third Formal Epistemology Festival in Toronto, the Philosophical Society in Oxford, the Pacific Division of the American Philosophical Association, my Brown-Blackwell Lectures at Brown University, and the Dot-fest in London. Thanks to David Chalmers, David Christensen, Stewart Cohen, Cian Dorr, Dorothy Edgington, Jane Friedman, John Hawthorne, Jeff Horty, Mark Kaplan, Maria Lasonen-Arnio, Jim Pryor, Josh Schecter, Susanna Siegel, MajaSpener, Ralph Wedgwood, Jonathan Weisberg, and Tim Williamson for helpful feedback; and special thanks to Lee Walters for comments which caused the intended final draft of this chapter to be the penultimate draft.

On the approach discussed here, then, defeasible reasons are mental states. But they are not just any kind of mental state; for the signature function of a defeasible reason is special, something unexecuted by most mental states. After all, the signature function of a defeasible reason is to generate *epistemic pressure* or *rational bias*. We begin with that working assumption and generalize it when necessary.

More specifically, we use the expression ' $xRB(\Phi)$ ' to mean that x is a reason to believe Φ , with the basic idea of a reason then being:

(R) $xRB(\Phi)$ =df. It is possible to become justified in believing Φ on the basis of x .¹

Here x is to be a mental state, the kind of thing on the basis of which one can rationally come to believe. We need not delineate, for present purposes, the exact kinds of mental states which play the x -role in (R). We need only assume that beliefs and perceptual states do so. They will be our focus in what follows.

Next we follow orthodoxy and recognize two kinds of reason: indefeasible and defeasible. The former generate epistemic pressure to believe, which cannot be wiped out or undone by the addition of information consistent with the information to hand. The latter generate epistemic pressure to believe which can be wiped out or undone by the addition of such information. Intuitively: indefeasible reasons are the meat and potatoes of rational deduction; and defeasible reasons are the meat and potatoes of rational induction, with reasoning of an inductive sort being the kind of reasoning which exploits the workings of *defeaters*. In turn these are mental states which wipe out or undo the rational bias put in place by reasons.

We use the expression ' $yD[xRB(\Phi)]$ ' to mean that y is a defeater for x as a reason to believe Φ , with the basic idea of a defeater then being:

(D) $yD[xRB(\Phi)]$ =df. (i) $xRB(\Phi)$,
& (ii) $\neg(x+y)RB(\Phi)$.

On this way of thinking, y is a defeater for x as a reason to believe Φ exactly when it is possible to become justified in believing Φ on the basis of x but not possible to become justified in believing Φ on the basis of x and y together.

Next we follow orthodoxy and recognize two kinds of defeater: rebutting and undercutting. Before explaining them, though, we flag an important assumption normally made about defeaters, an assumption which will come under attack later:

The Defeaters Assumption. Defeaters do their work because they are reasons to believe. They generate their distinctive kind of epistemic pressure—defeating epistemic pressure—in virtue of being reasons to believe.

This assumption plays a crucial role in standard thinking about undercutting and rebutting defeaters.

To see this, consider the latter. The basic idea of a rebutting defeater for x as a reason to believe Φ is:

(RD) $yRD[xRB(\Phi)]$ =df. (i) $xRB(\Phi)$,
& (ii) $yRB(\neg\Phi)$.

¹ See, for instance, Pollock (1987a: 35). Pollock emphasized defeasible reasons from the very beginning of his career. See his (1967), (1970), and his (1974), as well as the classics (1987a&b). I assume here that reasons under discussion are all of equal strength. Nothing turns on the assumption.

Or as it is normally put: y is a rebutting defeater for x as a reason to believe Φ exactly when x is a reason to believe Φ and y is a reason to believe $\neg\Phi$. Moreover: these conditions are meant to hold exactly when it is possible to become justified in believing Φ on the basis of x and possible to become justified in believing $\neg\Phi$ on the basis of y . But since rebutting defeaters are meant to be defeaters, of course, standard thinking about defeat—found at (D)—joins with standard thinking about rebutting defeat—found at (RD)—to entail that it is not possible to become justified in believing Φ on the basis of a reason to believe Φ plus a reason to believe $\neg\Phi$.

It is normal to deploy a similar line for undercutting defeaters. Unfortunately the details of this become delicate straightaway. We shall work up to them with a pair of vignettes: one involving the undercutting defeat of belief, the other involving the undercutting defeat of experience. The orthodox approach to undercutting defeat explains it in slightly different ways depending on whether the reason being undercut is itself a belief or an experience. The need for such a wrinkle in theory will not arise on the new view of undercutting defeat sketched later in this chapter.

Here is our first vignette:

The Polling Case

A pollster surveys 1000 voters in Texas at random, asking whether they will vote Republican or Democrat in the next election. Results generate belief in a claim about testimony:

T = 87% of respondents said they will vote Republican.

Belief in T then generates belief in a generalization:

G = Roughly 87% of Texans will vote Republican.

Suppose the pollster then comes to believe

U = Respondents decided their answer by coin flip.

Intuitively, belief in U is a defeater for belief in T as a reason to believe G. Yet belief in U is not a reason to believe $\neg G$. U is neutral concerning the truth-value of G. So belief in U is not a rebutting defeater for belief in T as a reason to believe G. Rather, belief in U is an undercutting defeater for that reason. As the discoverer of undercutting defeat, John Pollock, routinely put it in his work: belief in U ‘attacks the connection’ between belief in T and belief in G.² But how does belief in U play this role? How does it attack the epistemic link which manifestly exists between belief in T and belief in G?

Pollock answers this question by fleshing out a schema meant to slot into a schema for the undercutting defeat of belief:

$$\begin{aligned} B(U)UD[B(T)RB(G)] &=_{df.} \quad (i) \quad B(T)RB(G), \\ &\& \quad (ii) \quad B(U)RB(\neg X)]. \end{aligned}$$

The shape of this schema is familiar: belief in U is an undercutting defeater for belief in T as a reason to believe G exactly when belief in T is a reason to believe G and belief

² Pollock (1987a).

in U is a reason to believe some claim $\neg X$. This is the same theoretical shape used earlier in our approach to rebutting defeaters. And just as before the conditions put forward are meant to hold exactly when it is possible to become justified in believing G on the basis of belief in T and possible to become justified in believing $\neg X$ on the basis of belief in U. Since undercutting defeaters are meant to be defeaters, moreover, our general approach to defeat—found at (D) and endorsed by Pollock—joins with his schema for the undercutting of belief—found just above—to ensure that it is not possible to become justified in believing G on the basis of a reason to believe G plus a reason to believe $\neg X$.

This generates a simple question: what plays the X-role in the schema for the undercutting defeat of belief? What is the claim X such that belief in U's being a reason to believe $\neg X$ thereby makes belief in U attack the connection, in the way characteristic of undercutting belief, between belief in T and belief in G?

Throughout his work on undercutting defeat Pollock used surprisingly delicate forms of words to answer this question. And each time he took an initial stab at formulating such a form of words, he re-phrased it immediately. And often those re-phrasings did not look straightforwardly equivalent to what had been rephrased. In my view, this is diagnostic of a weak spot in Pollock's approach to undercutting defeat, with the weakness being exposed in a moment. For now, though, we shall flesh out the shape of Pollock's approach to undercutting defeat, so we can work effectively to its weaknesses.

In the second edition of *Contemporary Theories of Knowledge*, Pollock and Joe Cruz say that the undercutting defeater U in the Polling Case 'is a reason for doubting or denying that you would not have the inductive evidence unless G were true.'³ This loose form of words suggests a plausible thought in the Polling Case—and, as we'll see, a plausible thought in the analogue of the Polling case to do with visual reasons. In turn the plausible thought is this: any information to the effect that political facts on the ground were not responsible, in the Polling case, for the polling data in that case would *thereby* undercut belief in the polling data as a reason for thinking that G is true. In other words, any reason to deny T-because-G would itself undercut belief in T as a reason to believe G.

This plausible thought will make for an easy-to-understand and initially plausible theory of undercutting defeat. Unfortunately that theory is subject to clear counter-examples, as we'll see. And while the easy-to-understand view of undercutting defeat is not exactly faithful to Pollock's harder-to-understand approach, once we get clear on the counter-examples to the easy-to-understand theory it will be clear how to construct analogues for the harder-to-understand approach.

To begin, we create an easy-to-understand approach to the undercutting defeat of belief by appeal to the plausible idea sketched two paragraphs back. Let ' $\neg[\Phi \otimes \Psi]$ ' mean 'not-(Φ is true because Ψ is true)'. Then the rebutting-defeat-style schema for the undercutting of belief—the schema of the previous page—can be filled in as follows:

³ Pollock and Cruz (1987a: 196). I have renamed the claims in the example. Pollock used the same form of words in the singly-authored first edition of the book.

$$\begin{aligned}
 (\text{UDB-easy}) \quad B(\Delta)UD[B(\Phi)RB(\Psi)] &=_{\text{df.}} \quad (i) \quad B(\Phi)RB(\Psi), \\
 &\& \quad (ii) \quad B(\Delta)RB(\neg[\Phi \otimes \Psi]).
 \end{aligned}$$

In quasi-English: belief in Δ is an undercutting defeater for belief in Φ as a reason to believe Ψ exactly when belief in Φ is a reason to believe Ψ , and belief in Δ is a reason to believe not- $(\Phi$ is true because Ψ is true). These conditions are meant to hold exactly when it is possible to become justified in believing Ψ on the basis of belief in Φ , and possible to become justified—on the basis of belief in Δ —in believing not- $(\Phi$ is true because Ψ is true). And since undercutting defeaters are meant to be defeaters, of course, our approach to defeat—found at (D)—joins with the approach to undercutting defeat sketched above—to entail that it is not possible to become justified in believing Ψ on the basis of a reason to believe Ψ together with reason to believe not- $(\Phi$ is true because Ψ is true).

This easy-to-understand approach to the undercutting defeat of belief extends gracefully to the undercutting defeat of visual experience, our exemplar type of experiential reason. Consider another vignette:

The Visual Case

You have a visual experience representing a red object before you. On its basis you come to believe R: the claim that there is a red object before you. Then you are informed that

U = Local lighting is tricky in that it makes non-red objects look red.

Intuitively, belief in U is a defeater for your visual experience as a reason to believe R. Yet belief in U is not a reason to believe $\neg R$. U's truth is irrelevant to whether there is a red object before you. Belief in U is not a rebutting defeater for your visual experience as a reason to believe R. It is an undercutting defeater instead: somehow belief in U attacks the connection between your visual experience and your belief in R; and it does so in the way characteristic of undercutting defeat.

But how?

As we have seen, orthodoxy has it that the undercutting of belief turns on the content of one mental state—the belief which is the undercut reason—failing to be related aptly to the content of another belief—the belief which is formed on the basis of the undercut reason. This idea does not generalize to the undercutting of visual experience. The Visual Case (and vignettes like it) involve visual contents identical to belief-contents formed canonically on their basis. Yet no content is a reason for itself, much less an undercut reason; so unlike the undercutting of belief, the undercutting of visual experience cannot be entirely a matter of the content of mental states involved being aptly related to one another.

So how does it work?

The most natural answer—which dovetails with the easy-to-understand approach to undercutting above—turns on the fact that the existence of the defeasible reason itself, in the Visual Case, is explained by the very situation which makes true the belief formed on its basis. The relevant thought is that belief in Δ undercuts your visual experience of Φ as a reason to believe Φ when belief in Δ is reason to doubt or deny that you experience as of Φ because Φ . We can gracefully extend the easy-to-understand approach the undercutting defeat of belief, then, so that it covers visual reasons as well, by letting ' $\neg[V(\Phi) \oplus \Phi]$ ' mean 'not-(you visually experience as of Φ

because Φ '. Then we have an easy-to-understand approach to the undercutting defeat of visual reasons:

$$\begin{aligned} (\text{UDV-easy}) \quad B(\Delta)UD[V(\Phi)RB(\Phi)] &=_{\text{df.}} \quad (i) \quad V(\Phi)RB(\Phi), \\ &\& \quad (ii) \quad B(\Delta)RB(\neg[V(\Phi)\oplus\Phi]). \end{aligned}$$

In quasi-English: belief in Δ is an undercutting defeater for visual experience of Φ as a reason to believe Φ exactly when visual experience of Φ is a reason to believe Φ , and belief in Δ is a reason to believe that it is not the case that you visually experience as of Φ because Φ . These conditions are meant to hold exactly when it is possible to become justified in believing Φ on the basis of visual experience of Φ , and possible to become justified—on the basis of belief in Δ —in believing that it's not the case that you experience as of Φ because Φ . And since undercutting defeaters are meant to be defeaters, of course, the approach to defeat found at (D) joins with the approach to the undercutting defeat of visual experience found above to entail that it is not possible to become justified in believing Φ on the basis of a reason to believe Φ together with reason to believe it is not the case that you experience as of Φ because Φ .

This completes a picture of the traditional approach to reasons and defeaters. It is not strictly Pollock's approach to undercutting defeat, as we'll see. But it is decidedly in the spirit of Pollock's approach; and it has the considerable merit of being easy to understand. According to every approach in the neighborhood, including the easy-to-understand one we have sketched, reasons are mental states which generate epistemic pressure to believe; and defeaters succeed in destroying that pressure by generating their own epistemic pressure to believe.

This is very puzzling indeed. How can defeat itself spring from epistemic pressure to believe? In the next section, we flesh out this worry and unearth an assumption something like which must lie behind any approach to reasons and defeaters like Pollock's.

11.3 Combining Things

When we place our definitions of reason and defeat along side the Defeaters Assumption mentioned earlier—the view that defeaters do their work by being reasons to believe—puzzling explanatory schemata result. Suppose y is a rebutting defeater for x as a reason to believe Φ . The position before us entails the following explanation of rebutting defeat:

(ERD) It is not possible to become justified in believing Φ on the basis of $(x + y)$, despite it being possible to become justified in believing Φ on the basis of x alone, *because* it is possible to become justified in believing $\neg\Phi$ on the basis of y . In symbols:

$$\neg\diamond B(\Phi)_{[x+y]} \text{ even though } \diamond B(\Phi)_x \text{ *because* } \diamond B(\neg\Phi)_y.$$

This cannot be bedrock theory. There are stories of its form which do not work at all—e.g. when something intuitively irrelevant is substituted for $\neg\Phi$ —and for all we've been told y is a reason for $\neg\Phi$ only when sitting by itself in an agent's psychology, as the only relevant concern, whereas x is a reason for Φ when accompanied by further salient consideration. That sort of asymmetry has not been ruled out.

Or suppose y is an undercutting defeater for a belief x as a reason to believe Ψ . Then x will have a content Φ , and the overall position before us will entail the following explanation of the undercutting defeat of belief:

(EUDB) It is not possible to become justified in believing Ψ on the basis of $(x + y)$, despite it being possible to become justified in believing Ψ on the basis of x alone, *because* it is possible to become justified in believing $\neg[\Phi \otimes \Psi]$ on the basis of y [i.e. because it is possible to become justified in believing, on the basis of y , that not-(Φ is true because Ψ is true)]. In symbols:

$\neg \blacklozenge B(\Psi)_{[x+y]}$ even though $\blacklozenge B(\Psi)_x$ *because* $\blacklozenge B(\neg[\Phi \otimes \Psi])_y$.

Once again—and for exactly the same kind of reason—the story cannot be bedrock. There are stories of (EUDB)’s form which do not work at all—e.g. when something intuitively irrelevant is substituted for $\neg[\Phi \otimes \Psi]$ —and for all we’ve been told y is a reason for $\neg[\Phi \otimes \Psi]$ only when sitting by itself in an agent’s psychology, as the only relevant concern, whereas x is a reason for Φ when accompanied by further salient consideration. That sort of asymmetry has not been ruled out.

Or finally: suppose y is an undercutting defeater for visual experience v as a reason to believe Φ . Then v will have a content Φ —or so we are supposing—and the overall position before us will entail the following explanation of the undercutting defeat of visual experience:

(EUDV) It is not possible to become justified in believing Φ on the basis of $(v + y)$, despite it being possible to become justified in believing Φ on the basis of v alone, *because* it is possible to become justified in believing $\neg[v \oplus \Psi]$ on the basis of y [i.e. because it is possible to become justified in believing, on the basis of y , that not-(you experience as of Φ because Φ)]. In symbols:

$\neg \blacklozenge B(\Phi)_{[v+y]}$ even though $\blacklozenge B(\Phi)_v$ *because* $\blacklozenge B(\neg[v \oplus \Phi])_y$.

Once again the story does not look to be bedrock. Stories of its form do not work; and we have not been told how v and y work epistemically within an agent’s psychology.

Having said all that, *something* seems right about the stories about defeat just canvassed. The explanation of rebutting defeat—found at (ERD)—obviously has something going for it; and so do the explanations of the undercutting defeat of belief—found at (EUDB)—and the explanation of the undercutting defeat of visual experience—found at (EUDV). Why is that? More specifically, why do these stories seem to have something going for them despite the fact that they are obviously not bedrock theory?

Consider two answers.

Answer 1. When dealing with a composite state $(x + y)$, we presuppose that one can rationally believe a claim Φ on its basis, which can be rationally believed on the basis of x or y alone, *only if* it is rationally possible to believe all such Φ conjointly. When x is a reason to believe Φ and y is a reason to believe Ψ , therefore, the presupposition of Answer 1 is that composite state $(x + y)$ will be a reason to believe Φ , or reason to believe Ψ , only if it is rationally possible conjointly to believe Φ and believe Ψ .

This story renders cogent the explanation of rebutting defeat found at (ERD). After all, it is not rationally possible conjointly to believe Φ and $\neg\Phi$. When x is a reason to believe Φ and y is a reason to believe $\neg\Phi$, therefore, the composite state

$(x + y)$ will be composed of elements which are each reasons for claims it is not rationally possible conjointly to believe. Answer 1 thus makes sense of the explanation of rebutting defeat found at (ERD).

Unfortunately, the line does not handle the stories about undercutting defeat canvassed earlier [at (EUDB) and (EUDV)]. The key to Answer 1, after all, is this idea:

$(\Rightarrow A1)$ defeaters do their work because they are reasons to believe something which cannot be rationally believed while also believing the claim for which the defeated reason is a reason.

Suppose y is a rebutting defeater for x as a reason to believe Φ . Then y is a reason to believe $\neg\Phi$. Since it is not rationally possible to believe Φ and $\neg\Phi$, y is a reason to believe something which cannot be rationally believed while also believing the claim for which x is a reason. Hence y is a defeater for x (and vice versa).

Nothing like this occurs when a reason is undercut. To see this, recall the details of undercutting defeat. In the Polling Case, for instance, T is the claim that 87% of Texas respondents to a poll said they will vote Republican, G is the claim that roughly 87% of Texans will vote Republican, and U is the claim that respondents decided their answer by coin flip. When x is a belief in T and y is a belief in U , x is reason to believe G and y is an undercutting defeater for x as a reason to believe G . According to the story canvassed earlier, y undercuts x as a reason to believe G because y is itself reason to believe a complex negation: $\neg(T \otimes G)$. This is the claim that it is not the case that testimonial evidence is due to the fact that roughly 87% of Texans will vote Republican. The key thought behind Answer 1 will attempt to underwrite this story by saying that it is rationally impossible conjointly to believe G and $\neg(T \otimes G)$, i.e. it is not rationally possible conjointly to believe that roughly 87% of Texans will vote Republican and it is not the case that the testimonial evidence is due to the fact that roughly 87% of Texans will vote Republican.

That is simply not true. In the original Polling Case, after all, such a combination of beliefs is precisely what you should have if you come across further sound polling which indicates that G is true after all, that you were in a polling analogue of a Gettier case. This means that Answer 1 does not render cogent the overall story about the undercutting of belief that we have been considering.

The same holds true for the overall story about the undercutting defeat of visual experience we've been considering. To see this, recall the details of such undercutting. In the Visual Case R is the claim that there is a red object before you, and U is the claim that tricky lighting makes non-red objects look red. When v is a visual experience of R and y is a belief in U , v is a reason to believe R and y is an undercutting defeater for v as a reason to believe R . According to the line before us, y undercuts v as a reason to believe R because y is itself reason to believe a complex negation: $\neg(v \oplus R)$. This is the claim that it is not the case that you experience as of a red object before you because there is a red object before you. The key thought behind Answer 1 will attempt to underwrite this overall story by saying that it is rationally impossible conjointly to believe both R and $\neg(v \oplus R)$, i.e. it is rationally impossible conjointly to believe that there is a red object before you but that it is not the case that you experience as of a red object before you because there is a red object before you.

That is also not true. In the original Visual Case, after all, that is precisely what you should do if you come across further evidence that you suffer veridical hallucination. In that event you should believe precisely that there is a red object before you while denying, because of the tricky lighting, that it looks to you as if there is a red object before you because there is such an object before you. Answer 1 does not render cogent the story about undercutting of visual experience found at (EUDV).

Answer 2. When dealing with a composite state $(x + y)$, we presuppose that one can rationally believe a claim Φ on its basis, which can be rationally believed on the basis of x or y alone, *only if* it is rationally possible to believe all such Φ conjointly on the basis of $(x + y)$. When x is a reason to believe Φ and y is a reason to believe Ψ , therefore, the presupposition of Answer 2 is that composite state $(x + y)$ will be a reason to believe Φ , or reason to believe Ψ , only if it is rationally possible to believe Φ and to believe Ψ conjointly on the basis of $(x + y)$.

This story also renders cogent the explanation of rebutting defeat found at (ERD); and it does so, of course, for the same reason that Answer 1 did so. It is not rationally possible conjointly to believe Φ and $\neg\Phi$. Hence it is not rationally possible to do so on the basis of $(x + y)$. When x is a reason to believe Φ , and y is a reason to believe $\neg\Phi$, therefore, the composite state $(x + y)$ will be composed of elements which are each reasons for claims it is not rationally possible conjointly to believe. It follows that $(x + y)$ is composed of elements which are reasons for claims it is not rationally possible conjointly to believe on the basis of $(x + y)$. Answer 2 thus renders cogent the explanation of rebutting defeat found at (ERD).

Moreover, that Answer helps underwrite the explanations of undercutting defeat canvassed earlier. It has positive purchase not only on the story about undercutting defeat of belief—found at (EUDB)—but also on the story about undercutting defeat of visual experience—found at (EUDV). When x is belief in claim T—that 87% of Texas respondents said they will vote Republican, and y is belief in the claim U—that respondents decided their answer by coin flip—then, intuitively, x is a reason to believe G—the claim that roughly 87% of Texans will vote Republican—and y is an undercutting defeater for x . But as we have seen: it is rationally possible to believe that roughly 87% of Texans will vote Republican while also denying that the testimonial evidence turned out as it did because of the political facts on the ground. It is *not* rationally possible, however, to believe both these things *on the basis of* $(x + y)$. It is not rationally possible to believe—on the basis of a composite state consisting of a belief that 87% of respondents said they'd vote Republican and a belief that they did so by appeal to coin flips—that roughly 87% of Texans will vote Republican while also denying that 87% of respondents said that they would vote Republican because roughly 87% of Texans intended to vote Republican. Answer 2 renders cogent the explanation of undercutting of belief found at (EUDB).

It also renders cogent the explanation of undercutting of visual experience. When v is such an experience of R—the claim that there is a red object before you—and y is belief in the claim U—that tricky lighting makes non-red objects look red—then, intuitively, v is a reason to believe R and y is an undercutting defeater for v as a reason to believe R. But as we have seen: it is rationally possible to believe that there is a red object before you while denying that it looks to you as if there is because there is. It is *not* rationally possible, however, to believe both these things *on the basis of* $(v + y)$. It

is not rationally possible to believe—on the basis of a composite state consisting of a visual experience as of a red object before you, and a belief that tricky lighting makes non-red objects look red—that a red object is before you while denying that it looks to you as if a red object is before you because a red object is before you. Answer 2 renders cogent the explanation of undercutting of visual experience found at (EUDV).

The hope, then, is that the basic approach to reasons and defeaters before us turns out to be cogent if we assume that whenever one can rationally believe a claim Φ , on the basis of composite state $(x + y)$, which can also be rationally believed on the basis of a component of $(x + y)$, it is rationally possible to believe every such Φ conjointly on the basis $(x + y)$. This assumption would at least render sensible an aspect of the approach which looks decidedly puzzling at first, namely, the fact that on the approach defeaters function *as such* by virtue of being reasons to believe. One could be forgiven for not understanding this at first, for puzzling at the thought that being a reason to believe could itself invest a state with the power to wipe out another state's effectiveness as a reason to believe. Assumption 2 may be wrong, of course. But it helps to make sense of a very puzzling aspect of a Pollock-style approach to reasons and defeaters; and it does so for rebutting and undercutting defeaters alike.

Having said that, even if we accept Answer 2 we can make trouble for the general approach to undercutting defeat before us. Indeed with a bit of work we can see that undercutting defeaters turn out to function in an entirely different way than their rebutting cousins. Explaining why the Pollock-style approach to undercutting is no good will be our next task in the next section. Explaining the ways in which undercutting defeaters function differently than rebutting defeaters will be done in the section after that.

11.4 Counter-examples to Orthodoxy about Undercutting Defeaters

I begin with counter-examples to both directions of the easy-to-understand story about undercutting defeat. This will be done for the story's *spiel* about the undercutting of visual experience. It will be obvious, though, how to alter the examples so that they cover other forms of experiential reason and the undercutting defeat of belief. Once the easy-to-understand story has been found wanting, we'll turn our attention to Pollock's official story about undercutting defeat. By then it will be clear how to construct counter-examples to that story too.

Consider the following vignette:

The Milk Taster

Subject S tastes a bit of milk to see if it's gone off. Being a normal milk taster, S is unaware that her view of the milk is based on smell more than taste. Indeed we may suppose that S believes her view of the milk is not based on smell, not even in part, even though she knows, of course that she smells the milk as well as tastes it. When the milk taster imbibes the milk, however, she has a complex gustatory and

olfactory experience; and like other normal milk tasters she comes to believe on its basis that the milk is o.k. Like those milk-tasters, though, S is unaware that she bases her view of the milk on smell.

Suppose the milk taster is then told, by someone she trusts, that her nose is bunged up, that she is subject to random olfactory hallucination. This leads her, after a bit of reflection, to deny that her overall gustatory and olfactory experience of the milk was due to the milk being o.k. After all, she realizes that her overall gustatory and olfactory experience of the milk includes the olfactory part of that experience; and she believes herself to be subject to random olfactory hallucination. Nevertheless, her new information does not, and should not, lead her to change her view of the milk. She continues rationally to believe that the milk is o.k.; and she continues to do so on the basis of her complex gustatory and olfactory experience.

Let e be the milk taster's complex experience of the milk, OK be the claim that the milk is o.k., and H be the claim that the milk taster is subject to random olfactory hallucination. Then we have the following in the Milk Taster case:

- (i) e is a reason to believe OK: it is possible to become justified in believing OK on the basis of e .
- (ii) Belief in H is a reason to believe $\neg(e \oplus \text{OK})$: the milk taster can become justified in believing, on the basis of her belief in H, that it is not the case that her overall gustatory and olfactory experience of the milk is produced by the milk.
- (iii) Even after coming to believe H, the milk taster is rational in continuing to believe, on the basis of her overall gustatory and olfactory experience of milk, that the milk is o.k.

From these facts it follows that satisfying the easy-to-understand approach to the undercutting of experience is not actually sufficient for the undercutting of experience. With arrows marking the basing relation in thought, we can diagram the moral here as follows: when someone manifests this pattern of thought:

$$\begin{array}{ccc} B(\Phi) & & B[\exp(\Phi) \oplus \Phi] \\ \uparrow & & \uparrow \\ \exp(\Phi) & & B(U), \end{array}$$

the easy-to-understand approach to undercutting entails that U is an undercutting defeater for Φ -experience as a reason to believe Φ . That is simply not so. Milk tasters demonstrate that manifesting the cognitive situation diagrammed is insufficient for the undercutting defeat of experience. In turn that cuts against the easy-to-understand approach to such defeat.

Now consider another vignette

The Not-Because-Presupposer

Subject S is a normal person who begins with belief in the complex negation said by the easy-to-understand approach to be evidentially supported by undercutting defeaters. Let her start out with closed eyes, a desire to know if a red thing is before her, and a steadfast *presupposition* that it is not the case that the world will look as

it does, in situ, because the world is in fact the way it looks. S opens her eyes, it looks as if a red thing is before her, and S comes to believe on that basis that a red thing is before her.

She should not have formed the belief. S has undercutting defeat for her visual experience as a reason for her belief. But she does not have an undercutting defeater in the easy-to-understand sense; after all, she does not accept the negation of the complex because-claim in play here on the basis of a reason which supports that complex negation. She merely presupposes the negation is true, i.e. she merely presupposes that it is not the case that the world looks as it does, in situ, because the world is the way it looks. Contrary to the easy-to-understand approach, such a presupposition alone is sufficient for undercutting defeat.

Let v be the presupposer's visual experience as of a red object before her, R be the claim that a red object is before her, and $\neg(v \oplus R)$ be the complex negation in play. By stipulation S presupposes this complex negation. So we have the following in the Not-Because Presupposer case:

- (i) v is a reason to believe R : it is possible for the presupposer to become justified in believing R on the basis of v .
- (ii) The presupposer believes $\neg(v \oplus R)$ but not on the basis of a reason (or anything else). She merely presupposes that it is not the case that it looks as if R because R .
- (iii) The presupposer should not come to believe R on the basis of v : her belief that a red object is before her is irrationally formed on the basis of its looking as if a red object is before her.

From these facts it follows that satisfying the easy-to-understand approach to undercutting defeat is not necessary for such defeat. When someone manifests this pattern of thought:

$$\begin{array}{c} B(R) \\ \uparrow \\ v(R) \quad B[v(R) \oplus R], \end{array}$$

the easy-to-understand approach entails that $v(R)$ is an undefeated reason to believe R . That is not so. The Not-Because-Presupposer makes clear that manifesting the cognitive situation above is sufficient for undercutting defeat. In turn that cuts against the easy-to-understand approach to such defeat.

We may conclude that the easy-to-understand approach is no good. Its framing conditions are neither necessary nor sufficient for undercutting defeat. Milk Tasters show that those conditions can happen when undercutting defeat fails to occur. Not-Because-Presupposers show that undercutting defeat can happen when conditions put forth by the easy-to-understand view fail to occur. But that very view is not Pollock's, of course; it is a close cousin of Pollock's approach which is simply easier to understand. Perhaps the view's difficulties go away once complications are introduced, complications which make exactly for a harder-to-understand Pollock-style approach to undercutting defeat.

Unfortunately, difficulties faced by the easy-to-understand approach are close cousins of difficulties faced by Pollock's harder-to-understand view(s). To see this, recall how we arrived at the easy-to-understand theory. In the Polling Case: T is the claim that 87% of respondents said that they will vote Republican, G is the claim that roughly 87% of Texans will vote Republican, and U is the undercutting claim that respondents decided their answer by coin flip. Pollock and Cruz described U as 'a reason for doubting or denying that you would not have the inductive evidence unless G were true'. This led to the easy-to-understand idea that undercutting defeaters for belief are reasons for denying that one truth happens because of another, and the equally easy-to-understand idea that undercutting defeaters for experience are reasons for denying that experience which portrays the world a certain way occurs because the world is that way.

Just after making the remark which sets up these easy-to-understand ideas, Pollock and Cruz re-describe U as reason to believe that 'it is false that T would not be true unless G were true',⁴ and then they rephrase this new idea with the following gloss:

'This can be read more simply as 'T does not guarantee G'. Undercutting defeaters [of belief] are reasons for the claim that T does not guarantee G.'⁵

Further still, in his classic paper 'Defeasible Reasoning' Pollock describes undercutting defeaters as:

'reason for thinking it false that the premises of the inference [they attack] would be true unless the conclusion were true. More simply, we can think of [an undercutting defeater of belief] as giving us a reason for believing that (under present circumstances) the truth of the premises do not guarantee the truth of the conclusion.'⁶

These extra remarks on undercutting defeater complicate our discussion. They lead to a pair of hard-to-understand takes on the complex negation being reason for which is definitional, on Pollock's approach to undercutting defeat, of being an undercutting defeater. In turn this is true because the extra remarks lead to a pair of hard-to-understand takes on the claim negation of which plays the key role in Pollock's approach. As we are about to see neither of these hard-to-understand claims is equivalent to the other; and neither is equivalent to the easy-to-understand claim used earlier.

In the Polling Case, as we have seen, one claim the negation of which plausibly plays the key role is the claim that T is true because G is true (i.e. the claim that 87% of polled Texans said that they will vote Republican because roughly 87% of Texans will vote Republican). Earlier we symbolized the claim this way:

$[\Phi \otimes \Psi]$ Φ is true because Ψ is true.

The further remarks on the Polling Case quoted above generate two more claims which might play the key role in a Pollock-style theory of undercutting defeat. One is a subjunctive conditional we'll symbolize this way:

⁴ Pollock and Cruz (1987a: 197).

⁵ Pollock and Cruz (1987a).

⁶ Pollock (1987b: 486).

$[\Phi \rightarrow \Psi]$ Φ wouldn't be true unless Ψ were true.⁷

And the other is a claim about some kind of situational guarantee we'll symbolize this way:

$[\Phi_{pc} \gg \Psi]$ Φ 's truth guarantees Ψ 's truth, in present circumstances.

But notice: instances of any of these schemata do not entail analogue instances of the others. It might be the case that Φ is true because Ψ is true, for example, but not the case that Φ wouldn't be true unless Ψ were true, and also not the case that Φ 's truth guarantees Ψ 's truth in situ. Just think of a situation in which Φ is true because Ψ is true, but where Ψ concerns a weak back-up system in situ, one which happened quite improbably to have overridden the situation's dominant system Δ . Or it might be the case that Φ 's truth guarantees Ψ 's truth in situ, but neither the case that Φ is true because Ψ is true, nor the case that Φ wouldn't be true unless Ψ were true. Just think of a situation in which a weak back-up system Σ makes both Φ and Ψ true, while the dominant system Δ , the one overwhelmingly likely to have activated, only makes Φ true when activated. And so on. Instances of each of the indented schemata fail to entail analogue instances of the others.

Yet each of the schemata can be used in a Pollock-style approach to undercutting defeat. We have seen that the view which results when the easy-to-understand schema is used is itself no good. We are about to see that its problems have direct analogues for approaches based on the less-easy-to-understand schemata. Since the subjunctive-based one is by far the most common in Pollock's work on undercutting defeat, that will be our focus. Difficulties faced by the view of such defeat resulting from its use have direct analogues for the approach to undercutting defeat built from the other hard-to-understand schema.

Now, Pollock's work on undercutting defeat makes heavy use of subjunctive conditionals. His view of the undercutting of belief is normally put this way:

$$\begin{aligned} (\text{UDB-Pollock}) \quad B(\Delta)UD[B(\Phi)RB(\Psi)] &=_{\text{df.}} \quad (i) \quad B(\Phi)RB(\Psi), \\ &\& \quad (ii) \quad B(\Delta)RB(\neg[\Phi \rightarrow \psi]). \end{aligned}$$

Belief in Δ is said to be an undercutting defeater for belief in Φ as a reason to believe Ψ exactly when belief in Φ is a reason to believe Ψ , and belief in Δ is a reason to believe that it is not the case that Φ wouldn't be true unless Ψ were true. And each time Pollock offers a subjunctive-based approach to the undercutting defeat of belief he extends it directly to the undercutting of experience. With visual experience as our exemplar:

$$\begin{aligned} (\text{UDV-Pollock}) \quad B(\Delta)UD[B(\Phi)RB(\Phi)] &=_{\text{df.}} \quad (i) \quad V(\Phi)RB(\Phi), \\ &\& \quad (ii) \quad B(\Delta)RB(\neg[V(\Phi) \rightarrow \Phi]). \end{aligned}$$

Belief in Δ is said to be an undercutting defeater for visual experience as of Φ as a reason to believe Φ exactly when visual experience as of Φ is a reason to believe Ψ ,

⁷ I choose the dotted arrow here because it is quite clear from Pollock's career-long work on undercutting defeat that he thought some kind of binary connective was appropriate; but it is equally clear that he didn't have a stable conception of how to vocalize that connective in English.

and belief in Δ is a reason to believe that it is not the case that you wouldn't visually experience as of Φ unless Φ were true.

This approach to undercutting defeat will not work. Like its cousin the easy-to-understand theory, the conditions it lays down for undercutting defeat are neither necessary nor sufficient for such defeat. It is straightaway clear, for instance, that a presupposer counter-example plagues the view. Consider the following vignette:

The Not-Subjunctive-Presupposer

Subject S is a normal person who begins with belief in the complex negation said by (UDV-Pollock) to be evidentially supported by undercutting defeaters. Let her start out with closed eyes, a desire to know if a red thing is before her, and a steadfast presupposition that it is not the case that the world wouldn't look as it does if the world weren't the way that it looks. S opens her eyes, it looks to her as if a red thing is before her, and, on that basis, S comes to believe that a red thing is before her.

S should not have formed the belief. She has undercutting defeat for her visual experience as a reason for her belief. But S does not have an undercutting defeater in the style of (UDV-Pollock). After all, she does not accept the negation of the complex subjunctive on the basis of a reason supporting that negation. She presupposes the negation to be true. She presupposes that it is not the case that the world wouldn't look as it does if the world weren't the way that it looks. Contrary to (UDV-Pollock), this presupposition alone is sufficient for undercutting defeat.

Let v be S's visual experience as of a red object before her, R be the claim that a red object is before her, and $\neg[v \rightarrow R]$ be the complex negation in play. By stipulation S presupposes this negation. So in the Not-Subjunctive Presupposer case we have:

- (i) v is a reason to believe R: it is possible for the presupposer to become justified in believing R on the basis of v .
- (ii) S believes $\neg(v \rightarrow R)$ but not on the basis of a reason (or anything else). She presupposes that it is not the case that it wouldn't look as if R if it weren't the case that R.
- (iii) S should not come to believe R on the basis of v : her belief that a red object is before her is irrationally formed on the basis of its looking as if a red object is before her.

From these facts it follows that satisfying (UDV-Pollock) is not necessary for undercutting defeat. When someone manifests this pattern of thought:

$$\begin{array}{c} B(R) \\ \uparrow \\ v(R) \quad B[v(R) \rightarrow R], \end{array}$$

Pollock's theory entails that $v(R)$ is an undefeated reason to believe R. That is not so. The Not-Subjunctive-Presupposer makes clear that manifesting the cognitive situation above is sufficient for undercutting defeat. In turn this means that satisfaction of Pollock's conditions for such defeat is not actually required for it to occur.

Nor is the satisfaction of Pollock's conditions sufficient for undercutting defeat. Consider our final vignette:

The Mixed-up Assessor

Subject S is paid to watch a parade of Fs go by on a conveyor belt. The Fs come in two varieties: V1 and V2. S is paid to discriminate between them observationally. Like everyone, though, S is mixed-up about her capacity to do so. She thinks it is a purely visual capacity when in fact it is an olfactory one. S believes that when she is confronted with an F, somehow, she exploits subtle visual cues to register whether the F before her is V1 or V2. In fact S exploits subtle olfactory cues to make the discrimination. Sight plays no role at all.

S is aware, of course, that she smells Fs as well as sees them as they parade by on the belt. But S mistakenly believes—like everyone else—that Fs all smell the same to her while subtly looking different from one another. In fact the reverse is true: S discriminates V1s from V2s by exploiting subtle olfactory ways in which Fs are portrayed to be different from one another.

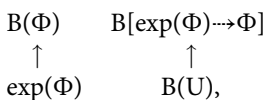
Suppose S is faced with an F. It looks and smells a certain way to her. In deploying her capacity to discriminate, S comes rationally to believe, on the basis of how the F smells, that it is V1 rather than V2.

Then S is told, by someone she trusts, that her nose is bunged up, that she is subject to random olfactory hallucination. This leads her to deny, after a bit of reflection, that the F wouldn't have smelled as it does if it weren't as it is. After all, she believes herself to be subject to random olfactory hallucination. Yet the information to that effect does not, and should not, lead her to change her view of the F. She continues rationally to believe that the F is V1 rather than V2; and she continues rationally to do so on the basis of how the F smells to her.

Let e be the mixed-up assessor's olfactory experience of the F, V1-F be the claim that the F before her is V1 rather than V2, and H be the claim that S is subject to random olfactory hallucination. Then we have the following in the Mixed-up Assessor case:

- (i) e is a reason to believe V1-F: it is possible to become justified in believing V1-F on the basis of e .
- (ii) Belief in H is a reason to believe $\neg(e \rightarrow V1-F)$: the assessor can become justified in believing, on the basis of her belief in H, that it is not the case that the F before her wouldn't have smelled as it did if it weren't that way.
- (iii) Even after coming to believe H, the assessor is rational in continuing to believe, on the basis of her olfactory experience of the F before her, that the F in question is V1 rather than V2.

From these facts it follows that satisfying Pollock's subjunctive-based approach to the undercutting of experience is not sufficient for such undercutting. With arrows marking the basing relation in thought, we can diagram the moral as follows: when someone manifests this pattern of thought:



Pollock's subjunctive-based approach entails that U is an undercutting defeater for Φ -experience as a reason to believe Φ . That is simply not so. Mixed-up assessors demonstrate that manifesting the cognitive situation diagrammed is insufficient for the undercutting defeat of experience.

Something is fundamentally wrong with Pollock-style approaches to undercutting defeat. Next we begin to make a case that such defeat works in an entirely different way than rebutting defeat. The key thought will be located at the heart of an unpublished thought experiment of Dorothy Edgington's.

11.5 Edgington's Burglar: Some Notation

The original thought experiment appears in an unpublished paper called 'Tale of a Bayesian Burglar'. That paper was sent privately by Edgington to David Lewis. It prompted Lewis to write an unpublished reply called 'Advice to a Bayesian Burglar'. As the titles make clear, the original discussion was framed within a Bayesian setting: epistemic agents were supposed to have point-valued degrees of belief, with the crucial issue at hand being how such degrees of belief should be updated.

In our discussion of undercutting defeat, however, we have proceeded under the assumption that belief is undercut, not degree of belief. That is standard practice in the literature, of course; but it is entirely inessential to the topic of undercutting defeat. After all, everyone (these days) agrees that experiential input can be undercut, no matter whether its initial impact is on belief or its degrees. This means that undercutting defeat requires theoretical treatment within the epistemology of degrees of belief just as it does within the epistemology of belief.

In what follows we prescind, therefore, from any take on the grain of states undercut by new information. We make no assumption about whether those states are coarse-grained like belief, disbelief or suspended judgement, or whether they are fine-grained like point-valued subjective probabilities. We distinguish simply between strong positive attitudes taken to a claim, neutral attitudes taken to a claim, and strong negative attitudes taken to a claim. And we represent them respectively as

- $\odot(\Phi)$ = strong pro attitude taken to Φ
- $\ominus(\Phi)$ = neutral attitude taken to Φ
- $\otimes(\Phi)$ = strong con attitude taken to Φ .

Now suppose you are a burglar. We set up our Burglar Case by appeal to three claims to which you take attitudes:

- A = There is an alarm on a given house.
- B = You break into the house.
- C = You get caught.

In the burglar case you begin with an initial take on things, then look at the house before you, and then update your view of the world. Hence you manifest two overall takes in the relevant thought experiment. Your initial take on the world is **old**(-); and your post-experience take is **new**(-).

Unsurprisingly, you want to break into the house but you do not want to get caught. You start with a strong pro attitude taken to the claim that you will get

caught, given you break into the house and it has an alarm. You start with a strong con attitude taken to the claim that you will get caught, given you break in and the house does not have an alarm. You know that you will look to see whether the house has an alarm, that you are next to certain to detect whether it does so, and that you will make up your mind whether to break in on that basis. You begin with a neutral take on whether the house has an alarm.

In the notation we shall be using the set-up is this:

Desire: $(B \rightarrow \neg C)$
 old- $\odot (C/BA)$
 old- $\odot (C/B \neg A)$
 $\odot (A)$.

But you are a sensible burglar. So the initial set-up makes for two further aspects of your initial take on things:

old- $\odot (B/A)$

and

old- $\odot (B/\neg A)$.

You begin with a strong con attitude to your breaking into the house given it has an alarm; and you begin with a strong pro attitude to your breaking into the house given that it has no alarm. Thus we have a key consequence of the initial burglar set-up:

old- $\odot (B/A) \neq$ old- $\odot (B/\neg A)$.

Next you look to see whether there is an alarm on the house. It visually appears as if there is an alarm on the house. This prompts a shift in your view. Your old neutral take on whether there is an alarm on the house—old- $\odot (A)$ —is replaced with to a new pro take on the claim that there is—new- $\odot (A)$.

Then something interesting should happen. You should go from a state in which your take on breaking in given there is an alarm is distinct from your take on breaking on given no alarm—as we have just seen—to a state in which the conditional commitments are identical. In other words: after updating on visual experience you should go from an old take on which:

old- $(B/A) \neq$ old- $(B/\neg A)$

to a new take on which:

new $(B/A) =$ new $(B/\neg A)$.

This means updating on your visual input makes your take on your breaking in go *from* being dependent on whether there is an alarm on the house *to* being independent of whether there is such an alarm.

Why is that? Well, as David Lewis put it:

‘[There is] a good reason why old- $\odot (B/A)$ and old- $\odot (B/\neg A)$ should differ: before you looked, you thought that if there was an alarm you would most likely spot it and be deterred, whereas if there wasn’t you would probably be undeterred and go ahead. But after you’ve already done your looking, and revised your [view], this reason no longer applies. There is no

good reason why $\text{New}(B/A)$ and $\text{New}(B/\neg A)$ should differ. Rather, they should be equal... for once you have finished looking the influence of the burglar alarm on whether you break in or not is over and done with.⁸

Updating on visual input shifts your take on breaking in from one on which it is dependent on whether there is an alarm to one on which it is not. This is so because you knew that facts about the alarm would control for your break-in behavior via your visual input, that mediating input had happened, and that your new take on there being an alarm was based on that input.

The lesson is both stark and general. Updating on visual episodes does not solely involve shifting commitment about the seen world. It also involves shifting commitment about the way in which first-order commitment about the seen world is itself founded. Visual updating routinely involves higher-order commitment about the basing of first-order commitment. When we observe the world and change our mind about it on the basis of that observation, we also routinely mark the fact that our first-order commitment is itself based on observation; and we do so with a higher-order commitment, another belief or investment of credence or some such. We register that our new view of the world is got by appeal to observation.

This happens so routinely that it has gone largely unnoticed in the literature. But it is crucial to the way that undercutting defeat works. There is a tight connection between when a commitment is undercut by new information and when there is a higher-order commitment about the way in which the undercut commitment is itself based. Explaining this is my next task.

11.6 Higher-order Mental States and Undercutting Defeat

Suppose you adopt a strong pro take on a claim Φ . Let U be the claim that source of information S is untrustworthy about Φ . Let $\Phi\text{-BOS}$ be any claim which forges a strong link between your take on Φ and source of information S .⁹ Then a bi-conditional of note will be true:

$$(\leftrightarrow) \quad \odot(U) \text{ undercuts } \odot(\Phi) \leftrightarrow (\exists x)x = \odot(\Phi\text{-BOS}).$$

In these circumstances a strong pro take on U undercuts such a take on Φ exactly when there is commitment to a strong link between one's take on Φ and source of information S .

Consider the right-to-left direction of $(\leftrightarrow)$:

$$(\text{RTL-}\leftrightarrow): \text{ Having } \odot(\Phi\text{-BOS}) \rightarrow \odot(U) \text{ undercuts } \odot(\Phi).$$

This looks plausible straightaway. If you are committed to the view that your strong pro take on Φ is itself based on source of information S , for instance, and then come

⁸ 'Advice to a Bayesian Burglar'. I have brought notation into line with this chapter.

⁹ $\Phi\text{-BOS}$ might be the claim that your take on Φ is based on S , or the claim that your take on Φ is likely to be based on S , or the claim that your take on Φ has a high objective chance of being based on S , etc. $\Phi\text{-BOS}$ can be any claim commitment to which amounts to commitment to there being a strong link between source of information S and your take on Φ .

to adopt a strong pro take on the view that S is untrustworthy about whether Φ is true, that strong pro take on S's untrustworthiness undercuts your strong pro take on Φ . And this looks to be so whether or not your old take on Φ was in fact based on source of information S.¹⁰

Similarly, consider the left-to-right direction of ($\leftrightarrow$):

(LTR- $\leftrightarrow$) $\odot(U)$ undercuts $\odot(\Phi) \rightarrow$ One has $\odot(\Phi - \text{BOS})$.

This also looks plausible. Suppose you are rationally committed to Φ but are not committed to there being a link between your commitment to Φ and source of information S. Then you come to adopt a strong pro attitude to the view that S is untrustworthy about whether Φ is true. Is your take on Φ undercut by your new commitment? The milk-taster and mixed-up assessor scenarios make it clear that the answer is No. If you do not have a strong pro attitude to the view that your take on Φ is linked to S, therefore, coming to adopt such an attitude to the view that S is untrustworthy about Φ will not undercut your take on Φ . The left-to-right direction of ($\leftrightarrow$) is just the contraposition of this thought.

The plausibility of each direction of ($\leftrightarrow$) suggests that we distinguish two fundamentally different types of defeat. The idea would be that rebutting defeaters generate a distinctive kind of pressure on their own—all by themselves, as it were—whereas undercutting defeaters generate a distinctive kind of pressure only in concert with higher-order commitment about the basing of lower-order commitment. That is a main suggestion of this chapter.

In the event, orthodoxy about rebutting defeat is fine. Rebutting defeaters do their work in splendid isolation, by being reasons to adopt a strong pro commitment of a certain sort. On the approach to undercutting defeaters suggested here, however, such defeaters work in an entirely different way; and that is why orthodoxy about them breaks down. Undercutting defeaters do their work in tandem with other mental states. They join forces with higher-order commitment about the basing of lower-order commitment.

Think of how dogmatists see the epistemic role of perceptual experience.¹¹ They say that visual states, for instance, generate epistemic pressure on their own to adopt a strong pro attitude to their contents, without appeal to background commitment or achievement. Non-dogmatists deny this, of course, insisting that visual states generate epistemic pressure only in concert with background commitment or achievement. My suggestion is that rebutting defeaters are the defeat-theoretic analogue of dogmatic experience, and undercutting defeaters are the defeat-theoretic analogue of non-dogmatic experience. Rebutting defeaters function the way dogmatists say perceptual experience works—in line with orthodoxy about rebutting defeat, without reliance on background epistemic factors; and undercutting defeaters function as non-dogmatists insist that perceptual experience works—contrary to orthodoxy

¹⁰ The same would be true for any other higher-order content commitment to which entails commitment to a strong link between S and your commitment to Φ .

¹¹ The term 'dogmatism' comes from Pryor (2000). Pollock was an early and vigorous defender of the view: see his (1974) or (1987), for instance.

about such defeat, in tandem with background commitment about the basing of lower-order commitment.

If this suggestion is right, the Defeater's Assumption is only partly correct. In our current terminology that assumption is this:

The Defeaters Assumption. Defeaters do their work because they are reasons to adopt a strong pro epistemic stance; they generate their distinctive kind of epistemic pressure—defeating epistemic pressure—in virtue of being reasons to do so.

My suggestion is that the Defeaters Assumption is true only for rebutting defeaters. Only they generate defeating pressure in virtue of being reasons to adopt a strong pro attitude. Undercutting defeaters do not work that way. They join forces with higher-order commitment about the basing of such attitude, with the resulting complex of mental states being no reason to believe at all. Rather, it is a reason to avoid basing commitment on untrustworthy sources of information. When an agent is committed to a strong link between her pro take on Φ and source of information S, then, and only then, does commitment to the Φ -untrustworthiness of S undercut her take on Φ . Rebutting and undercutting defeaters work in fundamentally different ways.

References

- Edgington, D. Unpublished. Tale of a Bayesian burglar.
 Lewis, D. Unpublished. Advice to a Bayesian burglar.
 Pollock, J. (1967) Criteria and our knowledge of the natural world, *Philosophical Review* 76(1): 28–60.
 Pollock, J. (1970) The structure of epistemic justification, *American Philosophical Quarterly*, Monograph series no. 4: 62–78.
 Pollock, J. (1974) *Knowledge and Justification*. Princeton, NJ: Princeton University Press.
 Pollock, J. (1987a) *Contemporary Theories of Knowledge*. Lanham, MD: Rowman & Littlefield.
 Second edition co-authored with Joe Cruz.
 Pollock, J. (1987b) Defeasible reasons, *Cognitive Science* 11(4): 445–80.
 Pryor, J. (2000) The skeptic and the dogmatist, *Nous* 34(4): 517–49.

Edgington on Possible Knowledge of Unknown Truth

Timothy Williamson

The chapter is a response to Dorothy Edgington's article 'Possible Knowledge of Unknown Truth' (*Synthese*, 2010), where she defends her diagnosis of the Church–Fitch refutation of the principle that all truths are knowable and analogous refutations of analogous principles, in response to my earlier criticisms of her diagnosis.* Using counterfactual conditionals, she reformulates the knowability principle and its analogues to withstand the Church–Fitch objection. In the present chapter, I argue that in order to avoid a kind of trivialization, Edgington needs to supply a more general constraint on how the knower is allowed to specify a counterfactual situation for the purposes of her reformulated principles, and that it is unclear how to do so. I also question the philosophical motivation for her reformulation strategy, with special reference to her application of it to Putnam's epistemic account of truth. In passing, I question how dangerous Church–Fitch arguments are for analogues of the knowability principle with non-factive evidential attitudes in place of knowledge. Finally, I raise a doubt about the compatibility of Edgington's reformulation strategy with her view that counterfactual conditionals lack truth-conditions.

* Dorothy Edgington and I were colleagues at Oxford from 2003 to 2006, while she held the Waynflete Chair of Metaphysics. I remember the joint graduate classes we gave in that period as some of the most enjoyable and rewarding teaching in which I have ever participated. Dorothy creates a relaxed, unthreatening, friendly atmosphere of straightforward intellectual co-operation, which (therefore rather than nevertheless) encourages everyone to aim for the highest standards of accuracy and clarity. In particular, she has a knack of finding the simplest, most perspicuous, least fancy example to make a point. These qualities come through in her writing too. She is a model for a way of doing philosophy that is deeply scientific but not in the least dehumanizing. This chapter originates in a talk given to the 2011 conference in honour of Dorothy at the Institute of Philosophy in London. The material was also presented to a class in Oxford. I thank both audiences, and in particular John Hawthorne and Jeremy Goodman, for useful comments. Special thanks to Lee Walters, who provided valuable detailed written comments on a draft of this chapter—and above all to Dorothy herself, for her wonderful contributions, both in person and in writing, both in teaching and in research, to philosophy.

12.1 Introduction

Many philosophers have been tempted by something like the idea that all truths are knowable. The idea is naturally formalized thus:

Knowability $\forall P(P \rightarrow \Diamond KP)$

Here $\Diamond$ and K stand for ‘it is metaphysically possible that’ and ‘at some time someone knows that’ respectively; $\rightarrow$ is just the material conditional, and the variable P takes sentence position.

Alonzo Church discovered the following refutation of Knowability. The principle entails its special case where the conjunction $P \& \neg KP$ is substituted for P . The consequent of the resulting instance is $\Diamond K(P \& \neg KP)$. But $K(P \& \neg KP)$ is impossible, because it entails both KP (since knowledge of a conjunction entails knowledge of its conjuncts) and $\neg KP$ (since knowledge entails truth). That refutes the consequent. Hence the special case reduces to $\neg(P \& \neg KP)$, the negation of its antecedent, and so to $P \rightarrow KP$. Thus Knowability entails the apparently stronger claim that every truth is known by someone at some time.¹ But that claim is silly. No one will ever know how many spiders were in my room exactly ten years ago. Therefore Knowability is false. As an anonymous referee for *The Journal of Symbolic Logic*, Church communicated his proof to Frederic Fitch, who was the first to publish it (Church, 1945; Fitch, 1963; Salerno, 2009b).

In 1985, Dorothy Edgington published a seminal reconsideration of the Church-Fitch argument (before its connection with Church emerged). While conceding that the argument refutes Knowability as formalized above, she explained a variant reading of the claim ‘Every truth is knowable’ that the argument does not refute. It depends on a distinction between the situation in which one knows and the situation one knows about. More specifically, her reading is this:

E-Knowability $\forall P \forall s ((\text{in } s : P) \rightarrow \exists s^* (\text{in } s^* : K(\text{in } s : P)))$

Here the variables s and s^* range over possible situations, possibilities that need not be maximally specific. E-Knowability says that whatever holds in a possible situation can be known to hold in that situation, but the knowing itself may take place in another possible situation. If, as before, one substitutes $P \& \neg KP$ for P , but in E-Knowability rather than Knowability, the consequent of the result says that in some possible situation s^* it is known that, in the possible situation s , P is an unknown truth. That involves no contradiction. Edgington’s 1985 paper is clearly one of the most original and thought-provoking treatments of the Church-Fitch arguments, and deservedly one of the most cited.

A couple of years later, I published a response to Edgington’s paper (Williamson, 1987b). My central objection was that since E-Knowability quantifies into an

¹ In intuitionistic logic, the argument yields only the weaker result $P \rightarrow \neg\neg KP$. This matters because some sympathizers with Knowability, such as Michael Dummett, were motivated by a form of verificationism that replaced classical with intuitionistic logic. Nevertheless, the argument presents significant difficulties even for such verificationists. I have discussed the argument in the intuitionistic setting (Williamson, 1982, 1988, 1992, 1994b); for a more recent treatment and further references see Murzi (2010). The present chapter assumes a classical setting, to which both Edgington and I are sympathetic.

epistemic context—the variable s occurs free in the scope of the operator K —we need some suitable constraint on how the knower in s^* is to specify the possible situation s , but none has been supplied. For one *can* quite naturally specify a situation by stating what is true in it ('the possibility that my ticket wins the lottery'), but if E-Knowability allows such specifications it becomes in a way trivial, since its consequent is verified by a possible situation in which someone knows the triviality that in the situation in which various things including that P are the case, it is the case that P .

Edgington also gave a variant of E-Knowability in notation more like that of Knowability, using modal operators including 'actually' (@), rigidly pointing back to the actual world, rather than quantification over non-specific situations:

$$\text{EA-Knowability } \forall P (@P \rightarrow \Diamond K @P)$$

EA-Knowability says that whatever holds in the actual world can be known in some world or other to hold in the actual world. It finesses the Church–Fitch objection to Knowability just as E-Knowability does, but is open to a similar objection. If a knower in a counterfactual world knows that in the actual world it is the case that P , how is she to specify the actual world? If she says 'the actual world' she specifies *her* world; but she was supposed to specify *our* world. If EA-Knowability allows the knower to specify the actual world by what is true in it, its consequent is verified by a possible world in which someone knows the triviality that in the world in which various things including that P are the case, it is the case that P , although in practice it may be impossible to specify all those various things in detail. Since Edgington treats human cognition as more concerned with unspecific possibilities than with maximally specific possible worlds, she concentrates on E-Knowability rather than EA-Knowability. I will do the same.

Edgington's 1985 paper is rich in examples. Her knowers specify an alternative possibility non-trivially by means of a counterfactual conditional: the possibility that would have obtained if such-and-such had been the case. The trouble is that providing a non-trivial verification of a claim does not show that the claim does not also have a trivial verification. Rather, one needs to clarify the claim so as to exclude trivial verifications. In the case of E-Knowability, that requires something like a general constraint on which ways of specifying an alternative situation are to count. Part of my critique involved showing that features of Edgington's examples that might seem promising candidates for extrapolation to a general constraint do not in fact generalize as required (Williamson, 1987a, 1987b, 2000: 290–301).

In 2010 Edgington published a long-awaited reply to my critique. Part of her reply is that her general approach does not commit her to those failed extrapolations, because she can treat my examples in other ways. I quite agree. But the point of my examples was to bring out ways in which her examples do less than one might think to adumbrate a general constraint of the sort required. Just by rejecting various incorrect generalizations of her cases, Edgington does not thereby provide a correct generalization of them.

Section 12.2 will discuss in more detail the problem of advancing from Edgington's discussion of her examples to a suitably general clarification of E-Knowability. In particular, it explains why just stipulating that the knowledge at issue must take some non-trivial form, without further specifying that form, does not solve the problem.

Of course, the clarification of E-Knowability should fit the intended philosophical point of proposing the principle. Section 12.3 questions Edgington's underlying motivation for her strategy of reformulating Knowability and similar principles, with special reference to her application of it to an epistemic account of truth once defended by Hilary Putnam. Section 12.4 briefly queries the relation between Edgington's general view of counterfactual conditionals and her use of them in defence of her reformulated principles.

12.2 Clarification of E-Knowability

Edgington makes clear that when she treats knowledge of a counterfactual as constituting knowledge *de re* about a possibility, the possibility she means is typically not just the possibility that so-and-so literally expressed by the antecedent, but rather a more specific possibility that would have obtained if so-and-so. For example, she says 'to have enough handle on *which* possibility one is talking about, one refers to it as the one that would have developed, had there been a course of history which diverged at a certain point from the actual history' (Edgington, 2010: 48). The courses of history she needs include something's being an unknown truth, but she does not write that into the antecedent of the counterfactual conditional itself, on pain of trivializing the knowledge at issue in E-Knowability.²

Some notation will facilitate the discussion. Edgington's possibilities are possible *situations*. We also allow impossible situations (although the quantifiers in E-Knowability do not range over them). A situation *s* is possible if and only if it is metaphysically possible for *s* to obtain. One way of specifying situations is by nominalizing sentences: the situation that *P* obtains if and only if *P*. A situation *s* *strictly implies* a situation *s'* if and only if it is metaphysically necessary that if *s* obtains then *s'* obtains. A situation *s'* is *less specific* than a situation *s* if and only if *s* strictly implies *s'* but *s'* does not strictly imply *s*. A possible *world* is a maximally specific possible situation, that is, a possible situation *s* no less specific than any possible situation. The locution 'in *s*: *P*' is equivalent to 'the situation *s* strictly implies the situation that *P*'. For present purposes, we may take 'situation' to apply to coarsely individuated items: situations are identical if and only if they are mutually strictly implying. In other words, $s = s'$ if and only if it is metaphysically necessary that *s* obtains if and only if *s'* obtains. We also assume that any situations have a conjunction, which (as a matter of metaphysical necessity) obtains if and only if all of them do. A situation *s* *counterfactually implies* a situation *s'* if and only if had *s* obtained, *s'* would have obtained.³ Since such counterfactual conditionals may be

² Edgington's take on her principle EA-Knowability is therefore radically at odds with that suggested by David Chalmers (2012, p. 31), which involves counterfactually specifying the actual world *w* by the infinite conjunction of all sentences in an imaginary canonical language true at *w*.

³ Could we have treated 'in *s*: *P*' as equivalent to 'the situation *s* counterfactually implies the situation that *P*' (rather than to 'the situation *s* strictly implies the situation that *P*')? That reading is uncharitable to Edgington, because it tends to undermine the difference between E-Knowability and Knowability. As an extreme case, if '*T*' is a tautology, Lewis's theory of counterfactuals makes 'the situation that *T* counterfactually implies the situation that *P*' equivalent to '*P*' itself, by his 'centering' axioms (6) and (7); Lewis, (1973: 132), which Edgington (2011: 83), accepts. Then E-Knowability entails Knowability, and thereby

contingent, which world we evaluate them with respect to makes a difference. For convenience, we take the default world of evaluation as a fixed possible world in which the putative knowing takes place. 'Counterfactually implies' should be understood accordingly.

One obvious objection to specifying a possibility as the one that would have obtained if so-and-so is that the definite description is improper. Many different situations would have obtained if so-and-so. In the notation just introduced, the situation that so-and-so counterfactually implies many different situations. For if a situation s counterfactually implies a situation s' , then s also counterfactually implies any situation less specific than s' . In particular, any situation counterfactually implies both itself and the trivial situation that necessarily obtains. Thus any non-trivial situation counterfactually implies at least two situations. Moreover, on most views of counterfactuals even the trivial situation counterfactually implies at least two situations, since it counterfactually implies both itself and a slightly non-trivial situation that obtains in all but some very remote possibilities.

The obvious fix is to specify a situation as 'the *most specific* situation counterfactually implied by s ', in other words, the situation counterfactually implied by s that strictly implies all situations counterfactually implied by s . Clearly, if s_1 and s_2 are both situations counterfactually implied by s that strictly imply all situations counterfactually implied by s , then s_1 strictly implies s_2 and vice versa, so $s_1 = s_2$ by the coarse-grained criterion of identity for situations above. Thus uniqueness is assured.

However, as so often, securing uniqueness jeopardizes existence. Why must there be a most specific situation counterfactually implied by the given situation s ? There is a most specific situation counterfactually implied by any given situation if and only if the counterfactual conditional commutes with conjunction, in the sense that a situation counterfactually implies some situations if and only if it counterfactually implies their conjunction. For suppose that the counterfactual conditional commutes with conjunction. Then s counterfactually implies the conjunction of all those situations it counterfactually implies separately. That conjunction is the most specific situation counterfactually implied by s . Conversely, suppose that there is a most specific situation counterfactually implied by any given situation. Let s be a situation, and s^+ the most specific situation counterfactually implied by s . Consider some situations S . Suppose that s counterfactually implies each one of S . Then s^+ strictly implies each one of S , and so strictly implies the conjunction of S (*strict* implication evidently commutes with conjunction), so s counterfactually implies the conjunction of S . Conversely, suppose that s counterfactually implies the conjunction of S . Then s^+ strictly implies the conjunction of S , and so strictly implies each one of S , so s counterfactually implies each one of S . Therefore the counterfactual conditional commutes with conjunction.⁴

succumbs to the Church–Fitch argument. Even on a weaker logic of counterfactuals and without such an extremely unspecific situation, related effects threaten. By contrast, 'the situation that T strictly implies the situation that P ' is equivalent to 'it is metaphysically necessary that P ', which yields no such collapse when substituted into E-Knowability.

⁴ Williamson (2007: 293–304), provides a suitable background logic of counterfactual conditionals and metaphysical modality for the argument in the text. Note the implicit use of the principle that if s_1

Although the commutativity of the counterfactual conditional with conjunction looks obvious, it is invalid in some mainstream logics of counterfactuals. In particular, David Lewis's preferred semantics for the counterfactual conditional makes it commute with finite conjunctions but not with infinite ones (Lewis, 1973: 19–21 and 132). In his terminology, the Limit Assumption may fail. Notoriously, Lewis allows cases such as this: for every positive length l , if this line had been longer then it would have been longer by less than l (for it would have been longer by at most $l/2$); but it is of course false that if this line had been longer then, for every positive length l , it would have been longer by less than l (for that is tantamount to saying that if it had been longer it would not have been longer).⁵ Thus for Lewis, even though this line could have been longer, there is no such thing as the *most specific* situation that would have obtained if it had been longer.

On a crude probabilistic account, a counterfactual conditional is true if and only if the chance of the consequent conditional on the antecedent exceeds a fixed threshold c less than 1. Then the counterfactual conditional fails to commute even with finite conjunctions. If the probabilities are doing much work, such an account may make it quite rare for there to be a most specific situation counterfactually implied by a given situation.

On Stalnaker's theory of conditionals (1968), every possibility counterfactually implies a unique possible world, which is the most specific situation (and a maximally specific possible situation) counterfactually implied by that possibility. This seems at odds with Edgington's emphasis on unspecific possibilities, although the contrast may be blurred by Stalnaker's qualification that it is often indeterminate *which* possible world a given possibility counterfactually implies (Stalnaker, 1984: 132–46).

Even if most possibilities do not counterfactually imply maximally specific possibilities, there may still be a most specific situation counterfactually implied by any given situation, for the counterfactual conditional may still commute with conjunction, as in the Lewis semantics with the Limit Assumption imposed. Commutativity will typically hold if the counterfactual conditional is a strict conditional restricted to contextually relevant worlds.⁶

counterfactually implies s_2 and s_2 strictly implies s_3 then s_1 counterfactually implies s_3 , which entails that an impossibility counterfactually implies anything (since an impossibility counterfactually implies itself and vacuously strictly implies anything). Although the vacuous truth of counterpossibles is somewhat controversial, see Williamson, 2007: 171–5), for a defence. Counterpossibles are in any case not very relevant to the purposes of Edgington's paper. The background logic has the principle that no possibility counterfactually implies an impossibility, so if we start in the realm of possibilities, neither strict nor counterfactual implication ever leads us outside that realm.

⁵ Perhaps there could have been lengths other than all actual lengths, if space had been differently structured, but we may assume that if the line had been longer space would still have been structured the same.

⁶ Lewis recognized such a view as an option but rejected it as defeatist (1973: 13). I write 'typically' rather than 'always' because the counterfactual conditional may fail to commute even with finite conjunctions on a dynamic semantics that evaluates it as a contextually restricted strict conditional at a context updated in response to the presence of the counterfactual itself in ways sensitive to its consequent. See also Gillies (2007).

That there always is a most specific situation counterfactually implied by a given situation is thus a contested assumption. Nevertheless, I will grant it to Edgington, because the commutativity of the counterfactual conditional with conjunction is a very plausible and attractive principle.⁷ Those sceptical of it may still grant it to Edgington for the sake of argument.

Let us return to the Edgingtonian proposal that knowledge of the counterfactual conditional that if A , C constitutes *de re* knowledge, of the most specific situation that would have obtained if A , that in it: C . Let s_A and s_C be the situations that A and that C respectively, and s_A^+ the most specific situation counterfactually implied by s_A (we have just granted that there is such a situation as s_A^+). In present notation, the proposal is that knowledge that s_A counterfactually implies s_C constitutes knowledge *de re*, of s_A^+ , that it strictly implies s_C .

The two contents of putative knowledge are at least guaranteed to have the same truth-value in the knower's world. By definition of s_A^+ , s_A counterfactually implies s_C in that world if and only if s_A^+ strictly implies s_C . It does not follow that the two contents strictly imply each other. For it may well be contingent whether s_A counterfactually implies s_C but non-contingent whether s_A^+ strictly implies s_C .⁸ The point is that the definite description used to fix the reference of the rigid situation term ' s_A^+ ' in the knower's world can pick out a different situation when used in another world. One may wonder how knowledge of the truth that s_A counterfactually implies s_C can constitute knowledge of the distinct truth *de re*, of s_A^+ , that it strictly implies s_C . But let us grant Edgington that, somehow, it can.

Once again, a kind of triviality threatens. For suppose that someone knows that s_A counterfactually implies s_C . Thus s_A does indeed counterfactually imply s_C . Let $s_{A \& C}$ be the situation that $A \& C$, and $s_{A \& C}^+$ the most specific situation counterfactually implied by $s_{A \& C}$. Then $s_{A \& C}^+ = s_A^+$. For in standard logics of the counterfactual conditional $\Box \rightarrow$, such as those of Lewis and Stalnaker, one can easily derive this theorem:

$$(*) \quad (A \Box \rightarrow C) \rightarrow ((A \Box \rightarrow P) \leftrightarrow ((A \& C) \Box \rightarrow P))$$

Informally: if C in the closest worlds in which A , then the closest worlds in which A are the closest worlds in which $A \& C$.⁹ Since s_A counterfactually implies s_C by hypothesis, s_A and $s_{A \& C}$ counterfactually imply exactly the same situations, by (*). Therefore, by definition of s_A^+ and $s_{A \& C}^+$, $s_{A \& C}^+ = s_A^+$. By assumption, knowledge that s_A counterfactually implies s_C constitutes knowledge *de re*, of s_A^+ , that it strictly implies s_C . By parity, knowledge that $s_{A \& C}$ counterfactually implies s_C constitutes knowledge *de re*, of $s_{A \& C}^+$, that it strictly implies s_C . Since $s_{A \& C}^+ = s_A^+$, knowledge *de re*, of $s_{A \& C}^+$, that it strictly implies s_C is knowledge *de re*, of s_A^+ , that it strictly implies

⁷ For a detailed defence of the commutativity principle see Fine (2012: 39–45). For discussion of its relation to the Barcan and Converse Barcan formulas see Williamson (2013: 127–9).

⁸ Strict implication is non-contingent in the modal logic S5, where everything is either necessarily necessary or necessarily not necessary.

⁹ Edgington seems to accept (*). In support of the centring principle, Edgington (2011: 83), says that '[g]ood reasons' for it are found in Walters (2009), where one of the two arguments for centering invokes (*) as a premise. Some authors have rejected (*); for a recent exchange see Ahmed (2011) and Walters (2011).

s_C (the use of Leibniz's law here is unproblematic because the context is *de re*). Therefore knowledge that $s_{A \& C}$ counterfactually implies s_C constitutes knowledge *de re*, of s_A^+ , that it strictly implies s_C . In effect, the very *de re* knowledge required to verify an instance of E-Knowability is constituted not only by the non-trivial knowledge that $A \Box \rightarrow C$, just as Edgington had in mind, but equally and independently by the utterly trivial knowledge that $(A \& C) \Box \rightarrow C$, a logical truth. Thus Edgington needs to gloss E-Knowability with some constraint on how the knowledge in question is constituted, in order to prevent this kind of trivialization of the principle. The constraint had better be reasonably general, in order not to invite the charge of ad hoc manoeuvring.

Edgington sometimes writes as though the problem were to distinguish between ways of specifying possibilities that achieve identifying reference and ways that fail to do so, where 'identifying reference' involves 'knowing which possibility one refers to' (Edgington, 2010: 49). But 'the most specific situation that would have obtained if $A \& C$ ' is in no way obviously worse than 'the most specific situation that would have obtained if A ' at letting one know which situation one is referring to. Indeed, the former description gives one *more* explicit information than the latter about what holds in the situation. The triviality is not intrinsic to the antecedent; it lies in the relation between the antecedent and the consequent.

Could Edgington cut the Gordian knot by simply explicitly requiring the *de re* knowledge of the situation to be constituted by non-trivial knowledge of a counterfactual conditional? The danger for this suggestion is of achieving the wrong sort of non-triviality. For example, let D be a lawlike scientific hypothesis, whose truth-value it is highly non-trivial (but possible) to determine; D has no special relevance to A or C . In fact, D is deeply nomologically impossible, whereas $A \& C$, though false, could very easily have been true. If the disjunction $(A \& C) \vee D$ had been true, its first disjunct would have been true.¹⁰ Thus the most specific situation that would have obtained if $(A \& C) \vee D$ is the most specific situation that would have obtained if $A \& C$. But for all one knows prior to scientific inquiry, D is true but $A \& C$ still false, in which case the most specific situation that would have obtained if $(A \& C) \vee D$ is the most specific situation that would have obtained if D (on Lewis's view, the knower's own world). Thus the very *de re* knowledge required to verify an instance of E-Knowability is constituted not only by the utterly trivial knowledge that $(A \& C) \Box \rightarrow C$ but also by the highly non-trivial knowledge that $((A \& C) \vee D) \Box \rightarrow C$. The trouble is that the non-triviality of the latter knowledge concerns only the falsification of D ; it has nothing to do with the relation between A and C . To allow such knowledge to verify the relevant instance of E-Knowability would entirely pervert the intended philosophical significance of the principle, as capturing the spirit of the claim 'All truths are knowable.' For the target truth in this

¹⁰ As Edgington reminds us, speakers often treat counterfactuals with disjunctive antecedents non-standardly, taking them to mean that each disjunct counterfactually implies the consequent, but she allows in a similar case that by heavy-handed wording we can enforce the intended standard compositional reading (Edgington, 2010: 47). That compositional reading is intended here. At a cost only in complexity, one could also replace the disjunction by its De Morgan equivalent in terms of conjunction and negation.

instance was *C*: we were to know *de re*, of a certain situation, that in it: *C*. But the core non-trivial knowledge we wound up with was of the falsity of *D*, which has no bearing on *C*. Knowledge of $((A \& C) \vee D) \Box \rightarrow C$ is trivial with respect to knowledge of *C* because it derives just from the outlandishness of *D*.

A similar point applies to the observation that E-Knowability and EA-Knowability are not entirely trivial because they imply the possibility of entertaining any truth of the form 'in *s*: *P*'. A realist may indeed allow the possibility of truths that cannot be so much as thought, let alone known. But a principle that rules out *that* possibility does not thereby capture the spirit of the claim 'All truths are knowable.'

One moral of the discussion so far is this. Without an underlying philosophical purpose, it is unwise to become involved in the project of fine-tuning something like Knowability in the hope of finding a principle in the vicinity that is neither trivially true nor trivially false, for the project lacks direction in the absence of a standard by which to judge whether a candidate principle meets the *point* of the original claim. Let us therefore examine Edgington's philosophical motivation.

12.3 Underlying Motivation for Edgington's Strategy of Reformulating Knowability

In *Knowledge and its Limits*, I complained that E-Knowability and EA-Knowability do not fit the arguments given by anti-realist philosophers such as Dummett for the claim that all truths are knowable, in which they reach verificationist conclusions by analysis of alleged conditions of understanding (Williamson, 2000: 300). Edgington concedes this point: 'Williamson may well be right that the sort of knowability I defend will be of little comfort to those who seek a systematic, verificationist theory of meaning.' She explains: 'I was not trying to defend knowability with that aim in mind. The holistic nature of evidential support—its strong dependence on background beliefs—makes such a project unfeasible, in my view' (Edgington, 2010: 51). I agree with her holistic objection to verificationist theories of meaning and understanding (it is not the only objection). But then why seek to refine Knowability?

Edgington states her general positive motivations for E-Knowability and EA-Knowability thus:

Rather, it struck me as implausible that hosts of very mundane facts should be in principle unknowable. Also, there are certain philosophical positions which, it seemed to me, would be defeated too readily by a Fitch-like argument, and which may be consistently restated by the technique I propose. (Edgington, 2010: 51)

I take the first reason first, the idea that it is 'implausible that hosts of very mundane facts should be in principle unknowable'. With no verificationism or anti-realism in the background, why should we *expect* even very mundane facts to be in principle knowable? The universe was not designed to facilitate our knowing. Perhaps what makes a fact 'very mundane' is that it is about objects, properties, and relations that are very familiar to us, which implies that we already have easy epistemic access to them. But it does not imply that we have easy epistemic access to every combination

of those objects, properties, and relations. Fitch–Church arguments concern facts that can be stated in very ordinary terms, but those terms are assembled into a subtle logical structure, involving universal quantification (over knowing subjects), negation, and an epistemic operator. Do we have any more reason to expect such a sentence not to state an unknowable fact than we have to expect the very mundane sentence ‘I spent my summer holiday in a village where one villager is a barber who shaves just those villagers who do not shave themselves’ not to state a contradiction? Roy Sorensen (1988) plausibly assimilates Church–Fitch unknowability to blind-spots, themselves rather mundane phenomena. Elsewhere I have argued that ordinary limits on our powers of perceptual and reflective discrimination surround us with clouds of unavoidable ignorance of very mundane fact, for reasons quite different from Church–Fitch arguments (Williamson, 1994a: 2000). For scientific purposes, it is pragmatically better for us not to give up too easily on the attempt to know, but we cannot require the universe to acknowledge our right to know even very mundane facts. Thus the weight of Edgington’s motivation needs to fall mainly on her second reason.

As an example of a philosophical position which would be defeated too readily by a Church–Fitch argument, and which may be consistently restated by her technique, Edgington offers an epistemic account of truth once defended by Putnam. I will examine this application of her technique in detail, as a test case.

Here is Putnam’s thesis, in Edgington’s words: ‘truth cannot transcend what could be predicted by a “theory” which is ideal by pragmatic standards’. She formulates a Church–Fitch argument against Putnam’s claim thus:

Putnam would concede that we may never obtain such an ideal theory. Suppose we do not. So there may be truths which are not predicted by any theory we ever devise. Let p be such a truth. So ‘ p and no theory ever devised predicts that p ’ is a truth. So by Putnam’s thesis, there is a possible ideal theory which predicts that: p and no theory ever devised predicts that p . But how *could* an ideal theory predict that p and no theory ever devised predicts that p ? On the reading generated by a parallel to Fitch’s argument, such a theory makes inconsistent predictions.

She then applies her technique to restate Putnam’s thesis:

But of course there is another, consistent reading: there is a possible, non actual theory which predicts that p and that none of the actually devised theories predicts that p .

(Edgington, 2010: 51)

An initial concern with Edgington’s formulation of the Church–Fitch argument is that it assumes that predicting anything of the form ‘ p and no theory ever devised predicts that p ’ constitutes making ‘inconsistent predictions’. But it does not constitute making *logically* inconsistent predictions. For the theory ‘No theory is ever devised’ is consistent by normal standards of logical consistency (it has models), even though devising it falsifies it, and it predicts (because it trivially entails) ‘No theory is ever devised and no theory ever devised predicts that no theory is ever devised’, which is of the form at issue. Perhaps Edgington intends a more pragmatic notion of inconsistency, such as being manifestly false-if-devised. A theory which predicts that p and no theory ever devised predicts that p is pragmatically inconsistent in that sense, and so clearly not ideal by pragmatic standards. Thus we may agree

with Edgington that the Church–Fitch argument refutes Putnam’s thesis on the reading the argument assumes.

To formulate the analogue of E-Knowability for Putnam’s thesis, we make the appropriate substitution for K in E-Knowability. Here is the result:

$$\begin{aligned} \text{E-Predictability} \quad & \forall P \forall s ((\text{in } s: P) \rightarrow \\ & \exists s^* (\text{in } s^*: \exists T (T \text{ is ideal} \ \& \ T \text{ predicts that in } s: P))) \end{aligned}$$

Here ‘ideal’ abbreviates ‘an ideal theory by pragmatic standards’. We may follow Edgington in assuming that a theory is ideal by pragmatic standards only if it is devised. For the analogue of the Church–Fitch argument for Putnam’s thesis, she uses the conjunction ‘ p and no theory ever devised predicts that p ’. If we substitute that conjunction for P and ‘the actual situation’ (on the required rigid reading of ‘actual’) for s in E-Predictability, and discharge the antecedent, we obtain this:

$$\begin{aligned} (!) \quad & \exists s^* (\text{in } s^*: \exists T (T \text{ is ideal} \ \& \ T \text{ predicts that in the actual situation:} \\ & (p \ \& \ \forall T^* (T^* \text{ is devised} \rightarrow \neg(T^* \text{ predicts that } p)))))) \end{aligned}$$

The reading Edgington suggests in the passage quoted above differs significantly from (!). She has ‘there is a possible, non actual theory which predicts that p and that none of the actually devised theories predicts that p ’, in which the first conjunct of the prediction, p , occurs outside the scope of ‘actually’. But the relevant instance of her principle cannot legitimately be read that way, for that instance results from substituting the whole conjunction ‘ p and no theory ever devised predicts that p ’ for the sentential variable P in the scope of ‘in s ’, so the first conjunct (p) is within the scope of ‘in s ’ just as much as the second conjunct is. Furthermore, she omits the qualification ‘ideal’ on ‘theory’, but without the conjunct ‘ T is ideal’ E-Predictability becomes much less interesting, since virtually anything is predicted by some non-ideal theory or other (for example, by an inconsistent theory). I will assume that Edgington was just writing loosely, and would accept E-Predictability as a fair version of what she intended, so that (!) is the proper consequent.

How does Edgington’s reformulation strategy fit the underlying philosophical motivation for Putnam’s thesis? His primary motivation for the thesis in the work she cites is an argument by *reductio ad absurdum* against the ‘metaphysical realist’ claim that a theory may be ideal by pragmatic standards yet false (Putnam, 1978: 125–6). For several reasons, Edgington’s choice of Putnam’s thesis to illustrate her strategy is unfortunate.

First, Putnam’s supposed *reductio ad absurdum* of metaphysical realism depends on dismissing in three lines any appeal to a causal theory of reference, because it would merely raise the question how the word ‘causes’ manages to secure unique reference (1978: 126). That has become notorious as the ‘just more theory’ move, and is generally, and rightly, regarded as illegitimate.¹¹ After all, one could similarly object to *any* account of reference (even a minimalist disquotational one) that it merely raises the question how the words in the account manage to refer. The objection depends on making utterly unreasonable demands on a theory of reference.

¹¹ Lewis (1984) gives an excellent detailed discussion of Putnam’s argument.

Since Putnam's underlying motivation for his thesis collapses anyway, it is unclear why we should be trying to reformulate that thesis. To illustrate the utility of her reformulation strategy, Edgington would do well to find a better-motivated philosophical thesis.

Second, having stated his argument, Putnam says that for it 'not to be just a new antinomy', 'one has to show that there is at least one intelligible position for which it does *not* arise' (1978: 127). For if the upshot of the argument is inconsistent with every intelligible position, then the argument is presumably fallacious. According to Putnam, however, the upshot of the argument is consistent with one intelligible position: a verificationist theory of understanding, on the model of Dummett's (1978: 127). He concludes that 'the theory of understanding has to be done in a verificationist way' (1978: 129). But the holistic grounds on which Edgington distances her strategy from verificationist theories of meaning tell just as strongly against verificationist theories of understanding. Thus Putnam's development of his thesis assimilates him to the very people to whom Edgington's strategy 'will be of little comfort', as we saw her already concede.

Third, Putnam's argument involves an appeal to, in effect, the upward and downward Löwenheim-Skolem theorems for the language of the ideal theory. Those theorems are standardly formulated for first-order non-modal languages, and fail for languages of many other sorts. Since E-Predictability states the ideal theory's predictions using the modal operator 'in *s*', defined above in terms of strict implication (although Edgington envisages us as being able to achieve a similar effect using counterfactual conditionals), delicate technical issues arise for the project of extending Putnam's argument to the richer language her reformulation requires. Thus Edgington's reformulation takes the language out of the class to which Putnam's own argument applies.

Fourth, E-Predictability is much too weak to capture Putnam's central claim against metaphysical realism, which is that ideal theories are not false. But nothing in E-Predictability requires the theory *T* not to be false in the situation *s** in which it is supposed to be ideal. Not even strengthening the conditional in E-Predictability to a biconditional would achieve that. Here is a toy illustration of the point. Assume that all formulas of the form 'in *s*: *P*' express either necessary truths or necessary falsehoods, since they are defined as strict implications. Suppose that ideal theories predict all necessary truths and no necessary falsehoods (perhaps in highly non-trivial ways), but may predict many contingent falsehoods. For example: in a sceptical scenario *Bad*, one is really a brain in a vat but appears to oneself to be in a non-sceptical scenario *Good*; the ideal theory in *Bad* predicts all and only those propositions that are true in *Good*. Then E-Predictability holds, as does its strengthening to a biconditional, but Putnam's central claim fails, and metaphysical realism is vindicated, because there are false ideal theories. Thus E-Predictability does not capture Putnam's central philosophical point.

If Edgington is to find a good illustration of the utility of her reformulation strategy, she will have to look further than Putnam's thesis.

At the end of her paper, Edgington follows Bernard Williams in emphasizing that one can imagine a scene without imagining oneself in that scene. As Williams notes, the point tells against Berkeley's reduction of the perceivable to the perceived, or of

the conceivable to the conceived (Williams, 1973b). Edgington suggests that in reducing the knowable to the known, the Church–Fitch argument makes a mistake similar to Berkeley’s (Edgington, 2010: 52). Williams’ point about the imagination is surely both correct and important. But it poses no threat to the use of the Church–Fitch pattern of argument against the claim that all truths are knowable, or similar claims. After all, what happens if we use that pattern of argument against the thesis that all truths are *imaginable*? Suppose that for some number n , there are exactly n stars and no one ever imagines that there are exactly n stars. Then, by the Church–Fitch reading of the imaginability thesis, it is possible that at some time someone imagines the conjunction that there are exactly n stars and no one ever imagines that there are exactly n stars. So what? The supposition that at some time someone imagines the conjunction that there are exactly n stars and no one ever imagines that there are exactly n stars is perfectly consistent, and possible. Someone could have imagined that conjunction. Of course, given that imagining a conjunction involves imagining each conjunct, the supposition entails that at some time someone imagines that there are exactly n stars. Thus the supposition also entails that the second conjunct of the imagined conjunction is false. Hence, unsurprisingly, the stronger supposition that the conjunction is both imagined and true *is* inconsistent. But what matters is that the Church–Fitch pattern of argument does *not* reduce the imaginable to the imagined. More specifically, it does not reduce the thesis that all truths are imaginable, on the reading amenable to that pattern of argument, to the thesis that all truths are imagined.¹² If applications of the Church–Fitch pattern of argument really involve some misunderstanding analogous to the one about the imagination that Williams diagnosed, one might expect the mistake to appear when one applies the pattern to the imagination itself: but none does. If anyone is in danger of committing Berkeley’s fallacy, it is the philosopher who feels tempted by an epistemic account of truth.

12.4 Edgington’s General View of Counterfactual Conditionals and Her Use of Them in Defence of Her Reformulated Principles

I will raise one more concern about Edgington’s treatment of the Church–Fitch argument. Edgington is best known for her important and innovative work on conditionals. In particular, she is the leading proponent of the view that they are not apt to be true or false, and should instead be evaluated in terms of conditional probabilities. She applies that view to subjunctive as well as indicative conditionals, in order to give a unified treatment of all conditionals (Edgington, 1995: 320–1 and 2004). How does her denial that counterfactual conditionals are truth-valued fit her use of them in her revision of the thesis that all truths are knowable and similar claims?

¹² Of course, the thesis that all truths are imaginable may well be false for quite different reasons, because not all propositions are capable of being entertained in thought, the negation of a proposition is capable of being entertained in thought only if the original proposition is capable of being entertained in thought, and one of the two propositions is true.

The question poses several problems. First, in her treatment of the original Church–Fitch argument, Edgington freely invokes knowledge-that with a counterfactual conditional content, which she requires to constitute *de re* knowledge of a possibility. But knowledge is supposed to be factive, to entail truth. How can one know that if this were the case then that would be the case if it is not *true* that if this were the case then that would be the case? A more general challenge is to explain what it means to know something with a probability-condition rather than a truth-condition. Many putative features of knowledge beyond factiveness are characterized in terms of the supposed truth-conditions of the objects of knowledge: for instance, reliability, sensitivity, and safety. They all lack obvious probabilistic analogues.¹³

Edgington herself has less sympathy for Knowability than for analogous epistemic constraints on truth in terms of non-factive evidential terms such as ‘predictable’ and ‘probable’ in place of ‘knowledge’.¹⁴ But even in those cases her reformulation strategy faces a related challenge: to explain how the role she postulates for counterfactual conditionals with respect to logically complex sentences such as E-Predictability relates to the semantics of those sentences.

In principle, the general problem arises independently of Edgington’s reformulation strategy. For example, the sentence ‘All those who would have posed a threat to the regime if they had been given the opportunity were rounded up and shot’ is presumably intelligible, but to whom does the plural noun phrase ‘those who would have posed a threat to the regime if they had been given the opportunity’ apply if the open sentence ‘*x* would have posed a threat to the regime if *x* had been given the opportunity’ lacks truth-conditions (relative to a context and an assignment of *a*

¹³ See Moss (2013) for a response to something like this challenge.

¹⁴ Edgington’s treatment of extensions of the original Church–Fitch argument to non-factive epistemic operators underestimates the extent of the difficulties in making the generalization (2010, p. 43). For example, let *E* be the operator ‘it is probable that’ (in an evidential sense), and consider the thesis that every truth can be probable, on the reading a standard Church–Fitch argument against that thesis assumes ($\forall P(P \rightarrow \Diamond EP)$). The argument requires the absurdity of $E(p \ \& \ \neg Ep)$, in the sense if not of logical inconsistency then at least of gross implausibility, which Edgington grants. But we can model $E(p \ \& \ \neg Ep)$ in epistemic logic within a framework for evidential probability as follows (see Williamson, 2000: 209–37 for background). For worlds, use the integers from $-n$ to $+n$, where n is odd, with a uniform prior probability distribution. The world k is epistemically accessible from the world j just in case $j - 1 \leq k \leq j + 1$ (this models a subject with limited powers of discrimination). The evidence at j is the set of worlds accessible from j (they are the worlds consistent with the evidence at j). The probability of *A* at j is the prior conditional probability of *A* on the evidence at j , in other words, the proportion of those worlds accessible from j where *A* holds. Let *EA* hold at j if and only if the probability of *A* at j is at least $2/3$. In the model, that means that *EA* holds at j if and only if *A* holds at two or three worlds accessible from j . Let *p* hold at all and only odd worlds. Then *Ep* holds at all and only even worlds, since the only odd world accessible from an odd world is itself, whereas two neighbouring odd worlds are accessible from an even world. Hence $p \ \& \ \neg Ep$ also holds at all and only odd worlds, so $E(p \ \& \ \neg Ep)$ holds at all and only even worlds, and in particular at 0. In other worlds, it is probable that: the world is odd and it is not probable that the world is odd. Of course, strictly speaking, what is needed for a Church–Fitch argument against the thesis that every truth can be *sometimes* probable is inconsistency in some sense (perhaps pragmatic) in the supposition that it can be *sometimes* probable that: *p* and it is *never* probable that *p*. But the only promise of such an inconsistency comes from the synchronic case. One may have to be content with a Church–Fitch argument against the stronger thesis that every truth can be both probable and true. Edgington herself makes that modification for an example involving eliminativism about folk psychology (2010: 43), but it seems to be needed much more widely. Although the unstrengthened claim ‘All truths can be probable’ lacks plausibility and motivation, its main problems do not stem from the Church–Fitch reading.

value to the variable x)? This is, of course, an instance of the classic Frege-Geach problem for non-truth-conditional accounts of sentences of some kind in terms of the alleged role of their unembedded occurrences (Geach, 1960, 1965). Since conditionals often embed somewhat awkwardly, the Frege-Geach problem might seem less pressing for them: one might hope to explain the cases in which they embed well as those in which some more or less *ad hoc* interpretative strategy is available. But Edgington's reformulation strategy treatment makes the Frege-Geach problem even more pressing for her, because on her view we grasp sentences of the form 'in s : P ' in effect as counterfactual conditionals, yet her reformulated principles embed such sentences in both the antecedent and consequent of a material conditional and in the scope of both universal and existential quantifiers.

Suppose that sentences of the form 'in s : P ' inherit probability-conditionality and non-truth-conditionality from counterfactual conditionals. Then Edgington's official reformulations embed probability-conditional but non-truth-conditional sentences in both the antecedent and consequent of a material conditional and in the scope of both universal and existential quantifiers, and Edgington needs to tell us what such embeddings mean. Presumably, she does not hope to do by constructing a systematic theory of meaning in terms of probability-conditions rather than truth-conditions, because that would be in effect to embark on the project of constructing a systematic, verificationist theory of meaning, in this case with a probabilistic form of verification. As we have already seen, on holistic grounds she regards such a project as 'unfeasible' (Edgington, 2010: 51). But how else is one to give a systematic, compositional semantics for such embeddings? Alternatively, if we interpret them non-compositionally, we are in effect reading the reformulated principles in some more or less *ad hoc* non-literal manner, which is an unhappy fate for what was supposed to be a canonical formulation of a significant philosophical doctrine.¹⁵

The non-truth-conditional treatment of 'in s : P ' is anyway dangerous for Edgington, because it is defined as above in terms of strict rather than counterfactual implication, so the alleged non-truth-conditionality is in effect being generalized from conditionals to modal operators, with a corresponding increase in its implausibility. In particular, when s is the (maximal) actual situation, 'in s ' is tantamount to the rigidifying 'actually' operator @, as in EA-Knowability, so presumably @ P is truth-conditional if and only if 'in s : P ' is too. To deny truth-conditions to sentences of the form @ P is particularly implausible, since we can quite easily and naturally specify truth-conditions for them.

The alternative is that sentences of the form 'in s : P ' and @ P have truth-conditions. In that case, Edgington's account requires our knowledge (or non-factive evidence) that their truth-conditions obtain to be somehow constituted by *de re* knowledge (or non-factive evidence) of a truth-conditionless counterfactual conditional, whatever such knowledge (or non-factive evidence) might be. If she could explain how such constitution is to work, the semantics of the reformulated principles would no longer

¹⁵ The failure of compositionality would be of a far more radical sort than that envisaged, for example, in Higginbotham, 1986: 33–7, where he argues that the semantic contribution of a conditional to a constituent is sensitive to features of the sentence in which it is embedded, but nevertheless in a systematic way.

pose a special difficulty for her, since it would still fall within the domain of truth-conditional semantics. But *how* could knowledge (or non-factive evidence) of something truth-conditionless constitute knowledge (or non-factive evidence) that a truth-condition obtains? One danger for Edgington in attempting to answer that question is that her answer might involve inadvertently supplying a plausible candidate truth-condition (perhaps a condition on probabilities) after all for the supposedly truth-conditionless thing—the counterfactual conditional—contrary to her view that it has none. Alternatively, reasons for not taking the explanation that way might turn out to undermine the constitution claim itself. But so far we lack even an attempt at an explanation.

In sum: reconciling Edgington's response to the Church–Fitch argument with her account of the semantics of conditionals is no easy matter, if it can be done at all.

12.5 Conclusion

In this chapter I have presented a number of serious difficulties for Edgington's defence of her treatment of Church–Fitch arguments. Nevertheless, I hope that my discussion makes it clear just how rich and rewarding are her two brief papers on the issue. There is surely much more to be said about all the problems I have raised here.

References

- Ahmed, A. (2011) Walters on conjunction conditionalization, *Proceedings of the Aristotelian Society*, 111: 113–20.
- Chalmers, D. (2012) *Constructing the World*. Oxford: Oxford University Press.
- Church, A. (1945) Referee reports on Fitch's 'A Definition of Value'. In J. Salerno (ed.), *New Essays on the Knowability Paradox*, Oxford: Oxford University Press, pp. 13–20.
- Dowe, P. and Noordhof, P. (eds) (2004) *Cause and Chance: Causation in an Indeterministic World*. London: Routledge.
- Edgington, D. (1985) The paradox of knowability, *Mind* 94: 557–68.
- Edgington, D. (1995) On conditionals, *Mind* 104: 235–329.
- Edgington, D. (2004) Counterfactuals and the benefit of hindsight. In P. Dowe and P. Noordhof (eds), *Cause and Chance: Causation in an Indeterministic World*, London: Routledge, pp. 12–27.
- Edgington, D. (2010) Possible knowledge of unknown truth, *Synthese*, 173: 41–52.
- Edgington, D. (2011) Conditionals, causation, and decision, *Analytic Philosophy* 52: 75–87.
- Fine, K. (2012) A difficulty for the possible worlds analysis of counterfactuals, *Synthese* 189: 29–57.
- Fitch, F. (1963) A logical analysis of some value concepts, *The Journal of Symbolic Logic* 28: 135–142; reprinted in J. Salerno (ed.), *New Essays on the Knowability Paradox*, Oxford: Oxford University Press, pp. 21–8.
- Geach, P. (1960) Ascriptivism, *The Philosophical Review* 69: 221–5.
- Geach, P. (1965) Assertion, *The Philosophical Review* 74: 449–65.
- Gillies, A. (2007) Counterfactual scorekeeping, *Linguistics and Philosophy* 30: 329–60.
- Higginbotham, J. (1986) Linguistic theory and Davidson's program in semantics. In E. LePore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, Oxford: Blackwell, pp. 29–48.

- LePore, E. (ed.) (1986) *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Oxford: Blackwell.
- Lewis, D. (1973) *Counterfactuals*. Oxford: Blackwell.
- Lewis, D. (1984) Putnam's paradox, *The Australasian Journal of Philosophy* 62: 221–36.
- Moss, S. (2013) Epistemology formalized, *The Philosophical Review*, 122: 1–43.
- Murzi, J. (2010) Knowability and bivalence: Intuitionistic solutions to the paradox of knowability, *Philosophical Studies* 149: 269–81.
- Putnam, H. (1978) *Meaning and the Moral Sciences*. London: Routledge & Kegan Paul.
- Rescher, N. (ed.) (1968) *Studies in Logical Theory*. Oxford: Blackwell.
- Salerno, J. (ed.) (2009a) *New Essays on the Knowability Paradox*, Oxford: Oxford University Press.
- Salerno, J. (2009b) Knowability noir: 1945–1963. In J. Salerno (ed.), *New Essays on the Knowability Paradox*, Oxford: Oxford University Press, pp. 29–48.
- Sorensen, R. (1988) *Blindspots*. Oxford: Clarendon Press.
- Stalnaker, R. (1968) A theory of conditionals. In N. Rescher (ed.), *Studies in Logical Theory*, Oxford: Blackwell, pp. 98–112.
- Stalnaker, R. (1984) *Inquiry*. Cambridge, MA: MIT Press.
- Walters, L. (2009) Morgenbesser's coin and counterfactuals with true components, *Proceedings of the Aristotelian Society* 109: 365–79.
- Walters, L. (2011) Reply to Ahmed, *Proceedings of the Aristotelian Society* 111: 123–33.
- Williams, B. (1973a) *Problems of the Self*. Cambridge: Cambridge University Press.
- Williams, B. (1973b) Imagination and the self. In B. Williams, *Problems of the Self*, Cambridge: Cambridge University Press, pp. 26–45.
- Williamson, T. (1982) Intuitionism disproved?, *Analysis* 42: 203–7.
- Williamson, T. (1987a) On knowledge of the unknowable, *Analysis* 47: 154–8.
- Williamson, T. (1987b) On the paradox of knowability, *Mind* 96: 256–61.
- Williamson, T. (1988) Knowability and constructivism, *The Philosophical Quarterly* 38: 422–32.
- Williamson, T. (1992) On intuitionistic modal epistemic logic, *Journal of Philosophical Logic* 21: 63–89.
- Williamson, T. (1994a) *Vagueness*. London: Routledge.
- Williamson, T. (1994b) Never say never, *Topoi* 13: 135–45.
- Williamson, T. (2000) *Knowledge and Its Limits*. Oxford: Oxford University Press.
- Williamson, T. (2007) *The Philosophy of Philosophy*. Oxford: Blackwell.
- Williamson, T. (2013) *Modal Logic as Metaphysics*. Oxford: Oxford University Press.

13

Prefaces, Sorites, and Guides to Reasoning

Rosanna Keefe

13.1 An Analogy

Is there an interesting relation between the Preface paradox and the Sorites paradox that might be used to illuminate either or both of those paradoxes and the phenomena of rationality and vagueness with which they, respectively, are bound up? In particular, if we consider the analogy alongside a familiar response to the Preface Paradox that employs degrees of belief, does this give any support to the thought that we should adopt some kind of degree-theoretic treatment of vagueness and the Sorites? I will argue that not only does it not give such support, but that some of the disanalogies count against such a treatment of vagueness more generally. Dorothy Edgington's work will be at the centre of my discussion. She considers analogies between the Sorites paradox and both the Preface and Lottery Paradoxes in her 1996 as part of the case for her important and original degree-theoretic account of vagueness.

I start by presenting the scenario within each paradox in such a way that brings out an analogy. First, a careful author believes each sentence in her book, but asserts in the Preface (acknowledging her fallibility) that some statement in her book is false.¹ Consider the argument from each of the accepted premises—the individual statements in the book—to their conjunction, or the quantification over them all. Although it is valid, this argument yields a conclusion the author explicitly rejects. She believes each of these premises, but the negation of their conclusion; moreover, she seems quite rational to do so. For the Sorites, consider a series of men, the x_i , starting with a 7 foot man and such that each is one hundredth of an inch shorter than the previous one. Consider x_k ; x_{k+1} is only one hundredth of an inch shorter, so surely if x_k is tall, then so is x_{k+1} , so we have a series of compelling conditionals, (C_1) 'if x_1 is tall then x_2 is tall', (C_2) 'if x_2 is tall then x_3 is tall' ... (C_i) 'if x_i is tall then x_{i+1} is

¹ See Makinson (1965) for the classic presentation of the Preface Paradox. Since I am primarily interested in the analogy with the Sorites paradox, I will not attempt to contribute to debates directly tackling the Preface Paradox, and will sometimes make simplifications in my discussion of it.

tall'. . . But the argument from the claim that x_1 is tall and these conditionals yields the absurd conclusion that x_{4000} is tall, though he is shorter than 4 foot. Such an argument can typically be constructed for any vague F on the basis of a suitable Sorites series of items. Both many-premised arguments concerning the Preface and Sorites start with many individually compelling statements and end up with a conclusion we reject (i.e. regarding the truth of all sentences in the author's book or the classification of the last, clearly non-tall member of the Sorites series).

In both cases the only rule needed is an utterly compelling, classically valid rule—conjunction introduction or modus ponens. And, in both cases, a natural informal first reaction to what has gone wrong has to do with the sheer quantity of premises. Whereas a move from two things that you believe to their conjunction is harmless, a move from thousands of individual things to the claim that they are *all* right is not safe. And whereas the move down a couple of steps of the Sorites series will not lead you astray, following hundreds of steps *will* do.

Both the Preface paradox and the Sorites paradox highlight an inconsistent set of claims, all of which a (normal, apparently rational) subject believes. In itself, this situation is not paradoxical as the subject can simply have got some things wrong and thus have some false beliefs. With the Preface, the *paradox* arises because we seem to be committed to certain principles about *rationality*, e.g. that it's never rational to believe inconsistent statements, or that it's always rational to believe the consequences of our beliefs. The key tension, then, is between the subject's apparently sensible beliefs and compelling principles about rationality. On the other hand, the Sorites is a paradox because we aren't interested just in a subject's *beliefs* but with the compelling hypothesis that each of the statements in the problematic set—the conditionals in the series, the claim that the first member of the series is F and the claim that the last is not- F —is *true*. This is an important difference between the two cases, but not yet a reason to discard the analogy.

There is another disanalogy that will, I argue, turn out to be particularly significant. In the Preface case, our author believes *each* of the statements but not *all* of them conjoined (i.e. the generalization saying they are all true). But in the Sorites case, the subject typically believes each of the conditionals *and* believes all of them, or the generalization over them taken together, namely 'for all i , if x_i is tall then x_{i+1} is tall'. Indeed, one of the most popular formulations of the Sorites paradox starts from a variant on that very generalization, e.g. with the premise:

(C_g) 'If x is tall and y is one hundredth of an inch shorter than x , then y is also tall.'

Again, adding the premise that x_1 is tall, we can derive the absurd conclusion that x_{4000} is tall.

By contrast, there is no way of compressing the Preface into a short, two-premise argument whose premises are still plausible. Any appeal to the analogy will have to downplay this popular version of the Sorites paradox and perhaps explain away the intuition that the generalized premise is true. Edgington, for example, claims that the 'long Sorites' (the one with many individual conditional premises) is 'basic', in contrast with the 'short Sorites' (containing a single generalized or inductive Sorites premise) (1996: 311). We will come back to this below.

13.2 Degrees of Belief and the Preface

An appealing response to the Preface Paradox involves recognizing degrees of belief: our author should (and does) believe each of the individual statements in the book to a degree just less than certainty and the slight doubts can add up, explaining her rational disbelief in the large conjunction of them all.² The degrees of belief can be modelled probabilistically, and compelling principles of rationality can then be formulated to replace a non-probabilistic requirement such as 'believe the conclusion of a valid argument if you believe the premises'.³ A familiar result from Ernest Adams shows how the probability of a conclusion of a valid argument relates to the probabilities of its premises. If we call 1 minus the probability of p the improbability of p , we can say that the improbability of the conclusion cannot exceed the sum of the improbabilities of the premises.⁴ This allows a valid argument with 10 premises of probability 0.9 to have a conclusion with probability 0. Rational degrees of belief can be assumed to follow this structure, so that your degree of disbelief in the conclusion of a valid argument should not exceed the sum of the degrees of disbelief you have in the premises. Our author can then remain rational: there are so many statements in the book that the slight uncertainties (i.e. degrees of disbelief) in each add up to at least 1.

Acknowledging degrees of belief in a Preface-type case can help guide us with how to reason in the face of uncertainty: by recognizing the extent to which our beliefs fall short of certainty and the number of premises, we can exercise the appropriate level of caution in drawing our conclusion, in line with the Adams principle. So with a short argument with just a couple of premises, we can remain confident in the conclusion, whereas lots more premises prompts lots more caution. Could appeal to a similar degree-theoretic framework in a theory of vagueness provide a similar story about how to reason in the face of vagueness?

Many theorists have sought to approach vagueness through a degree-theoretic framework whereby, for example, tallness can come in degrees, and the degree of tallness can drop through a Sorites series from the degree 1 definite cases to the degree 0 definitely not-tall cases, with different borderline cases taking different degrees in between. Many such theories employ degrees of truth: they regard truth as coming in degrees between 0 and 1 and the degree assigned to 'x is tall' is taken to gradually drop from 1 (or very nearly 1) to 0 (or very nearly 0) through our Sorites series of men of decreasing height. Can some such theory provide the best guidance for our reasoning when vagueness is involved?

² See, e.g., Foley (1992). My initial description of the paradox made no reference to degrees of belief, but we can use the above thought to explain our author's rationality, at least if we assume a correspondence between (non-quantitative) rational belief and rational belief above some threshold of degree of belief; cf. Foley's 'Lockean thesis'. See, e.g., Hawthorne and Bovens (1999) on locating the threshold, which can change between people or across contexts.

³ Or, perhaps, 'don't believe both the premises and the negation of the conclusion of a valid argument': the differences between different options here will not matter for our purposes. See Field, 2009: 259) for one option of such a probabilistic principle (where $P(A)$ is the subject's degree of belief in A): '(D*) If it's obvious that $A_1, \dots, A_n$ together entail B , then one ought to impose the constraint that $P(B)$ is to be at least $P(A_1) + \dots + P(A_n) - (n - 1)$ in any circumstance where $A_1 \dots A_n$ and B are in question.'

⁴ Adams (1966). See also Edgington (1996).

In Edgington's presentation of reasoning with uncertainty and vagueness, she discusses two analogous 'scare stories' (1996: 300, 302): the Preface suggests that we should only trust reasoning that starts from premises of which we are certain and the Sorites suggests that we can only trust valid arguments when our premises are completely clearly true. Given the vast extent of both uncertainty and vagueness that we deal with, this threatens to leave us very frequently without trustworthy arguments to use. Her approach builds on the thought that an appeal to degrees gets us out of these two problems and allows us to avoid the Scare Story. I will argue that it does not help with vagueness.

13.3 Degree-Theoretic Accounts of Vagueness: Edgington's Option

On Edgington's powerful degree theory, the structure of degrees of belief discussed above in relation to the Preface paradox is mirrored by another degree theoretic structure that she labels 'verities', or 'degrees of closeness to clear truth' (rather than the degrees of *truth* that other theorists employ).⁵ The degrees assigned to complex statements thus exhibit a probabilistic structure. This, again, is unlike other degree theoretic accounts of vagueness according to which the degree of truth of a compound sentence is a function of the degrees of its components—for example, where the degree assigned to a conjunction is the minimum of the degrees of the conjuncts. On a probabilistic model, the degree of the conjunction depends on the relations between the conjuncts, just as the *probability* of a conjunction is not just determined by the probability of the conjuncts. More specifically, Edgington employs the idea of 'conditional verity', where 'the conditional verity of B given A is the value to be assigned to B on the hypothetical decision to count A as definitely true' (1996: 306). For example, if Tim and Tek are both borderline tall but Tim is slightly taller than Tek, then the conditional verity of 'Tim is tall given that Tek is tall' should be 1. For if we were to take it as true that Tek was tall then we would have to count Tim as tall too, since he is taller. But if we consider 'Tek is tall given that Tim is tall', the verity will be less than one: deciding that Tim is tall doesn't yet settle that the shorter Tek is also tall. Rather, Edgington maintains, the decision induces a rescaling of the borderline cases shorter than Tim, rendering Tek near the top of the resulting borderline cases, so that 'Tek is tall' would then have value close to 1 and so this is the value to be assigned to 'Tek is tall given that Tim is tall'.⁶

The value of compound sentences is then calculated using this notion of conditional verities, with the exception of negation, where the verity of not-A is 1 minus the verity of A. The verity of 'A&B' is the value of A multiplied by the conditional verity of B given A. So, for example, 'Tek is tall and Tim is tall' has the same verity as 'Tek is tall' (since the conditional verity of 'Tim is tall given Tek is tall' is 1). But 'Tek

⁵ See Edgington (1996) for her theory. Cook (2002: 239) interprets Edgington's verities as degrees of truth, but she avoids this commitment to degrees of truth (1996: 299). I will not focus on this issue here; see my 2012, p. 462 for some further discussion.

⁶ See Keefe (2000: 99–100) for worries about the notion of conditional verity, at least for cases other than straightforward ones like the first described above.

is tall and 'Tim is *not* tall' appropriately gets verity 0, since 'Tim is not tall given Tek is tall' gets assigned 0 (if we were to decide that Tek was tall, the taller Tim would also count as tall and 'Tim is not tall' would get verity 0).⁷ The natural interpretation of the conditional 'If A then B' takes it as equal to the conditional verity of B given A.⁸ The individual conditional premises of the Sorites—the (C_i) —will then each have a verity close to 1, for the verity of Fx_{i+1} given Fx_i will be close to 1 (as we saw with 'Tek is tall given Tim is tall').

As Edgington demonstrates, Adams's result has a parallel within her structure of verities: the unverity (1 minus the verity) of the conclusion of a valid argument cannot exceed the sum of the unverities of the premises. She calls this the 'verity constraining property of valid arguments' (1996: 307). The application to the Sorites then shows how the many premises that are each highly plausible, and indeed have a verity close to one, can yield a false conclusion, despite the validity of the argument. Verity seeps away, given the many premises falling short of verity 1, resulting in a false conclusion despite the validity of the argument.⁹

Just as the appeal to degrees of belief can show how an author can rationally believe all the sentences in their book and yet believe that there is something false in it, appeal to verities can show why we find each of the Sorites conditionals—the (C_i) —compelling, because they have verity of nearly 1, even though they entail something false. The analogy may seem to help understand or support the theory of vagueness in question here, but I will argue that the position is not as appealing as it may seem.

Let us return to Edgington's 'Scare Story' and ask how far it helps to introduce verities and the verity constraining property of valid arguments. The worry with the Preface was that it shows that we can be justifiably highly confident of each of a group of claims (those in the body of the book), without that giving us grounds for confidence in a consequence of them (the claim that they are all true), suggesting that we should only trust reasoning that starts from premises of which we are certain. The story about degrees of belief, coupled with the probability-constraining property, ensures that we have guidance beyond the cases when our premises are certain. The parallel Scare Story in the vagueness case would suggest that we can only trust valid arguments when our premises have verity 1; the verity-constraining property promises a way out. The conclusions of our valid arguments cannot drop below certain

⁷ This illustrates an advantage of Edgington's model over the kind of degree-theoretic alternatives discussed in the next section. According to the latter, 'Tek is tall and Tim is not-tall' will get the minimum value of 'Tek is tall' and 'Tim is not tall' which will be around 0.5, even though we are strongly inclined to regard that statement as false. I won't examine this aspect of the debate here, however, focusing instead on issues more directly related to the analogy between the Preface paradox and the Sorites.

⁸ See Edgington (1992: 202). Edgington (1996: 307) resists commitment to this identification (which is problematic in combination with some of her other views on conditionals), setting aside the question of whether it is correct. We will continue with the simplification of assuming it is correct for the purposes of considering the Sorites paradox. An alternative – which would follow her 1996 more closely – would have been to concentrate on the Sorites formulated with the material conditional, $A \supset B$ defined as $\neg(A \& \neg B)$, e.g. with the series of compelling premises $\neg(Fx_i \& \neg Fx_{i+1})$.

⁹ In my 2012 I express a worry about this perspective on the Sorites given Edgington's denial that verities are degrees of truth: 'if it isn't truth that seeps away (as on the more standard interpretation of degrees of truth), why does the conclusion end up false?' (p. 462). I put that aside here.

levels, though the more premises there are with verities less than 1, the lower the conclusion can drop.

This may, then, help with many-premise Sorites where the reasoning goes wrong because it uses too many premises of verity close to but less than 1. But the 'short Sorites' cannot be solved the same way: (C_g), 'if x is tall then anyone one hundredth of an inch shorter is also tall' is assigned verity 0 and thus counts as clearly false (see Edgington, 1996: 311). It cannot be that this seems compelling because it is nearly true, so this counts against the theory. Edgington sees the 'long Sorites' as 'the basic form': 'the universally quantified statement in the short version has to be understood via the relatively basic, relatively less complex statements which are its grounds' (1996: 312), but she does not say much to support this claim. If this commits us to a corresponding explanation of our beliefs about the respective premises, then this prompts various psychological questions that we cannot address here. But, the account of our beliefs it suggests is not appealing. It is not that we survey all the instances of (C_i) and believe the generalization over them all on that basis: we have not formed beliefs about most of the instances of (C_i) independently (I have never considered Fx_{237} , for example). Rather, for any arbitrary instance, our reasons for believing it are general ones, e.g. that one hundredth of an inch is too little difference in height to make the difference to whether someone is tall. And that kind of reason surely supports (C_g) directly. It is more plausible to think we believe that generalization and (perhaps implicitly) believe each instance on those grounds.

Does Edgington's framework provide a guide to reasoning and one that answers the Scare Story? More specifically, insofar as it provides a guide, is it successful in such a way as to lend support to Edgington's theory and the employment of a degree-theoretic structure in the semantics of vagueness? A semantic theory can always yield a guide to reasoning in the sense that it tells you the status of a conclusion given the status of premises and thus we should be guided by the expectations about the status of the conclusion that this delivers, given the status of the premise. We want to know if it delivers a *good* guide: what reasons do we have for thinking that its verdicts on what counts as good reasoning are right? And considering the guide in this way will only help one's actual reasoning if we have access to what the semantic statuses of our premises are.

In the case of the framework surrounding the Preface, introspection (in judging one's own level of confidence in something) is a really good guide to one's degree of belief: even if we can't assign an *exact* degree this way, we are reliably accurate to a great extent in such judgements. This allows us to follow the instructions on what attitude to take to our conclusion given those degrees of belief in the premises. Things are not so straightforward with verities, however. In some cases, our degree of belief in a proposition is a good guide to its verity. In particular, it may be typically roughly right with atomic predications: the verity of ' x_i is F ' drops through the Sorites series in tandem with our inclination to classify x_i as F . But that is not enough to help with the arguments and reasoning we are interested in, for they will rarely involve atomic cases alone: the kinds of reasoning we are interested in all involve compound sentences. With non-atomic sentences, the extent to which we are inclined to believe something can be a bad indication of its verity, even when we're in possession of the key facts and would judge the verity of the relevant atomic sentences roughly

correctly. For example, as we've seen, (C_g), the Sorites inductive premise, has verity 0, despite our strong inclination to believe it.

This might suggest that you should always calculate the appropriate degree-theoretic status for the premises by using the semantics: consider what the framework assigns to a sentence of that structure given the verities of the components and certain relations between them. But that would require you to know the status of the components, and we can't always assume that when we reason from a complex sentence, we do have access to that information. For example, this would be no good for standard employment of elimination rules (where the idea is to reason from complex sentences to the components, so shouldn't depend on prior knowledge of those components). Contrast this with the uncertainty case: our author's low degree of belief in 'every statement in my book is true' is not based on a calculation of what that degree must be given various degrees of beliefs in the instances.

Note, that this isn't simply saying that the problem is that we're sometimes wrong in our assessment of the premises and that we can find a premise compelling when it is in fact a long way from nearly clearly true and quickly leads to a false conclusion. Any guide to reasoning must allow for this: it tells you how to extend your beliefs and/or whether to draw a given conclusion *given* what you believe and can't guard against error in your beliefs about the premises. In the treatment of the Preface, we must, of course, acknowledge that the author has a false belief in one of the claims in her book. But the degree-theoretic treatment of uncertainty doesn't just tell her that, which would provide no guidance when she has no reason to reject any one rather than another. It tells her, more helpfully, to lower her degree of belief in all the premises. Similarly, responding to the contradiction in the Sorites set-up by simply telling the subject that a premise is false provides little guidance. Edgington does not do this for the long Sorites (mirroring the probabilistic uncertainty case), but does do that for the short Sorites.

With Edgington's theory we could maintain that we *should* have a low degree of belief in the Sorites inductive premise because the semantics dictates that. But if there's no independent support of the pattern of degrees it yields, then the availability of the picture gives no support for the use of degrees in the theory of vagueness. When degrees are employed in the Preface case, this shows how the typical (compelling) combination of degrees of belief in the Preface propositions is rational after all. It predicts the degree of belief in the claim that everything in the book is true and that makes it a story that fits well with our beliefs and so a plausible story about them. The story with verities isn't like that. A story that assigns degrees that don't correspond to our actual degrees of belief needn't be any better off than a view like supervaluationism, according to which (C_g) is false.

The degree theorist may say that we *have* to recognize that the main premise of the short Sorites is false—so has verity 0—given the truth of the other premise and the falsity of the conclusion. But the degrees would then be doing no work in that story. Note that we have no reason to introduce degrees into our semantics if they are just used in an explanation that would be available to us by considering uncertainty alone. For example, it might be said that in advising how one should respond to the long Sorites, we should tell the subject to have a lower degree of belief in each of those conditionals, rather than presenting them with a specific premise that they should

reject. But this advice can be combined with any theory (e.g. an epistemic theory of vagueness) without assuming a role for verities or degrees of truth in the semantics or in the treatment of vagueness: we can just appeal to the account of reasoning with uncertainty that we've already accepted. There's no reason to think that the degrees of belief assigned on this basis correspond to anything relevant to the *semantics* of vagueness and so this constitutes no support for a degree theory of vagueness.

The analogy between the Preface Paradox and the Sorites Paradox thus does not support theories that require degrees in their semantics. And we cannot sustain the appealing degree-theoretic thought that something seeming true indicates that it is at least nearly true when, e.g., a sentence like (C_g) seems true but has verity 0. This is unlike in the Preface-type case where if we think a premise is true but we recognize we are less than certain of it, we can modify our confidence in the conclusion to an extent determined by the number of uncertain premises. That was how we avoided the original version of the Scare Story. We have seen that Edgington's story about vagueness does not deliver an equally satisfying solution to the corresponding Scare Story. Cases like the short Sorites sustain the threat that we should abandon reasoning with vague premises that seem true, because their seeming true is not a good indication that they are *nearly* true (or that we should think they are) and they can thus lead to a false conclusion in very few steps.

13.4 Non-probabilistic Alternatives and the Scare Story Again

The probabilistic structure of Edgington's degree theoretic account of vagueness is atypical of such accounts of vagueness. Many other accounts have been proposed and defended according to which the degree assigned to a complex sentence is a function of the degrees assigned to the components.¹⁰ Different options are available corresponding to different choices of the function taken to capture each of the connectives. We will express 'the degree of A ' as $|A|$. The usual definition of conjunction has $|A \& B| = \text{Min}\{|A|, |B|\}$, though on some accounts it is defined as the product of $|A|$ and $|B|$. Most commonly, $|A \vee B| = \text{Max}\{|A|, |B|\}$ and $|\sim A| = 1 - |A|$. There is more variation over the definition of the conditional, which can affect the evaluation of the key premises of the long Sorites, the (C_i) , where often the value of the conditional will just drop below 1 to the degree that the consequent is less true than the antecedent.¹¹

An account of the quantifiers will also be needed to evaluate the short Sorites, with a popular option for the existential quantifier following on from the most common definition of disjunction so such that the existential generalization is as true as its most true instance. This may seem to offer us a way to preserve the thought that sentences (like the premises of the Sorites) can seem true because they are nearly true, reinstating the parallel with the Preface paradox. For (C_g) 'if x is tall then anyone one

¹⁰ See e.g. Keefe (2000: ch. 4) for a summary and discussion of this range of theories and Smith, 2008 for a recent book-length defence of one option.

¹¹ E.g. according to Machina (1976), $|A \supset B| = 1$ when $|A| \leq |B|$ and $= (1 - |A|) + |B|$ otherwise.

hundredth of an inch shorter is also tall' can come out nearly true because the generalization takes the value of the least true instance and each of the (C_i) are nearly true.¹² One worry, though, is that such a degree theory of vagueness is not compatible with a good answer to Edgington's Scare Story and, in particular, it cannot provide an explanation of why we don't just need to abandon all our usual reasoning once we acknowledge uncertainty/vagueness. To explore this in more detail, we need to consider the various different definitions of validity that can be (and have been) adopted in degree-theoretic frameworks. For, starting from the standard classical definition of validity as necessary preservation of truth, there are several different ways to generalize this to accommodate degrees of truth.

On one popular degree-theoretic story about validity, an argument is valid if and only if its conclusion cannot be less true than its least true premise.¹³ This is a sense in which degree of truth can count as being preserved in valid arguments. On such an account, all Sorites arguments will be invalid. For example, the least true of the (C_i) (the individual conditional premises) will be nearly true on most accounts and yet the conclusion will be (at least almost completely) false. Moreover, modus ponens will be invalid too, as, for example, if $|A| = 0.5$ and $|B| = 0.2$, then on the most common account of the conditional, $|A \supset B| = 0.7$ and the other premise, A , has value 0.5, yet the conclusion, B , drops to value 0.2 (i.e. below the value of either premise). The Scare Story is then particularly pressing if we are obliged to give up such fundamental rules of inference on acknowledging the possibility of vagueness (and sentences less than degree 1 true). Edgington says that this treatment of modus ponens (and other classically valid reasoning) 'licenses modus ponens on clearly true premises, and licenses nothing on not-clearly-true premises. . . it leaves us inferentially impotent in the presence of vagueness' (1996: 303).

Even on an alternative account of the conditional according to which some of the (C_i) are not nearly true, those conditionals will never be as false as the conclusion. For example, Smith's account of the conditional (2008: 268) employs the classical equivalence between $(A \supset B)$ and $(\sim A \vee B)$, thereby allowing some of the (C_i) to drop to around 0.5 (e.g. for the cases around the middle of the borderline cases when both $\sim Fx_i$ and Fx_{i+1} have values around 0.5). But none of those conditionals will drop much *below* 0.5, and so not as low as the conclusion.¹⁴ So, this does not allow us to preserve validity within the Sorites according to the popular definition of validity above. As I'll reiterate below, we also lose the explanation that the conditionals seem true because they are nearly true.

An alternative definition of validity in a degree-theoretic framework identifies it with preservation of degree 1.¹⁵ This can be seen to generalize the classical definition of validity (necessary preservation of truth) in a different way, regarding degree 1 as truth and only requiring preservation of that. On this account, the Sorites argument

¹² This, for example, will be the verdict on the account in Machina, 1976, which is also the one Williamson argues that the degree theorist should accept (Williamson, 1994, pp. 114–18).

¹³ E.g. Machina, 1976; Forbes, 1983.

¹⁴ If Fx_{i+1} is much below 0.5, so will Fx_i be, in which case $\sim Fx_i$, and so the conditional (C_i) , will have a value higher than 0.5.

¹⁵ E.g. Peacocke, 1981.

comes out valid. But we get no answer to the Scare Story, for the account does not help us reason with premises of values less than 1. Again, we threaten to be left ‘inferentially impotent in the presence of vagueness’ (Edgington, 1996, p. 303).

Nicholas J. J. Smith advocates an account of validity that is different again and which he claims does preserve classical logic. According to that account, an argument is valid if and only if, whenever the premises are >0.5 degrees true, the conclusion is ≥ 0.5 .¹⁶ It might then be hoped that, unlike with the earlier definitions of validity, we get guidance for premises that have degrees <1 without abandoning the likes of *modus ponens*. Is *this* kind of account better equipped to deal with the Scare Story? I argue not. For one, the guidance it gives you as to how to reason will need to assume that what you care about is whether the conclusion is ≥ 0.5 . That does not seem to reflect our reasoning practices. One kind of scenario we may well want to rule out in our reasoning is the kind where we start from a couple of highly compelling premises that are very nearly true and draw a conclusion that is barely over half true. Yet the Smith definition of validity allows some such reasoning to be valid.

How does Smith deal with Sorites arguments: how can they deliver a false conclusion, despite compelling premises? On Smith’s account of conditionals, one of the premises (C_i), is just less than 0.5 degrees, for the pair Fx_i and Fx_{i+1} whose values straddle the 0.5 boundary, will be such that $\sim Fx_i$ and Fx_{i+1} are both <0.5 . So, the argument can be valid despite the conclusion being ≤ 0.5 , since one of the premises is <0.5 . Similarly, the generalization (C_g) comes out less than 0.5 true. These cases help show that Smith’s account is another account on which we cannot explain that the main premise of the Sorites is so compelling by saying that this is because it is nearly true. Indeed, this explanation cannot even be adopted for some of the individual conditionals (so the situation is even more serious than for Edgington). As with Edgington’s account, the semantics does not have the same practical use for our reasoning, given that we cannot directly assess the right degree to be assigned to our premises in the way we can judge uncertainty of complex sentences. The subject’s appeal to her degrees of belief (roughly, the extent to which she find something compelling) corresponds to degrees of truth at best only in atomic sentences, such as with the individual predications Fx_i through the Sorites series. But our *reasoning* requires complex sentences such as the conditionals in the Sorites argument. This is unlike the case where we introduce degrees of uncertainty to confront the Preface Paradox, where our uncertainty in simple or complex premises can typically be taken at face value and fed into the guidance of how uncertain we should be in our conclusion.

So the analogy with the appealing degree-theoretic approach to the Preface Paradox breaks down in the vagueness case. None of the various choices on the definition of validity or the account of the connectives will give us what we need. One shortcoming has built on the initial observation of the disanalogy between the short Sorites and the corresponding compressed version of the Preface paradox. Once we recognize that the short Sorites is compelling, a tension is generated. On the one hand we want to acknowledge that its main premise is highly compelling, so should come out with a high degree value, otherwise we cannot maintain that ‘seems true is nearly

¹⁶ Smith, 2008, p. 222.

true'. And if we abandon that, we fall foul of Edgington's Scare Story again, since the vagueness of compelling premises does, after all, seem to give us a reason to abandon reasoning with them; this is unlike in the uncertainty case where apparent near certainty in premises does allow us to carry out short inferences and where there is a clear story about how our confidence should decline in longer cases. On the other hand, if the generalized premise of the short Sorites does get a high degree value, the shortness of the short Sorites makes it a case where premises of high values can lead you dramatically astray, so yet again we face the Scare Story, this time because it shows we can't even trust short arguments with premises that are nearly true.

This undermines the analogy between the Sorites and the Preface and undercuts what might have looked like a good reason to pursue a degree theory of vagueness, while also serving as a good illustration of why such theories can't fulfil the roles hoped of them. This should not, of course, be taken as an attempted refutation of degree theories of vagueness. It is still open to degree theorists to accept the way degrees of truth or verities diverge from degrees of belief and maintain that some choice of definitions of the connectives and of validity provides the right story about truth-values and logical consequence and can guide our reasoning. We need, they may go on, to accept those definitions and work with them rather than following our intuitions about, for example, the degrees of complex sentences. At the very least, some of the apparent advantages of this kind of approach will be lost, however, including the appealing analogy with reasoning with uncertainty.

In giving a semantics, Edgington and other degree theorists provide some kind of guide to reasoning: we should reason as the semantics dictates, e.g. assuming the conclusions of an argument to take values as dictated by the semantics and the values of the premises. A guide of this form can be good or bad (according to whether the semantics is good or bad), so its mere availability says nothing in favour of the semantics in question. One hope for the analogy with the Preface could have been that this analogy provides support for the chosen semantics for vagueness. But this is undermined by the fact that those semantics for vagueness do not have the same support from the agreement with our intuitions as that displayed in the uncertainty case, as demonstrated by considering generalisations like (C_g).

In the next section I turn to a theory of vagueness that I claim provides us a better guide to reasoning. The analogy with the Preface Paradox plays no role here, but I claim that this, in fact, works to the theory's advantage. We do not need to employ a degree theoretic semantics to reason with vague expressions; we can have a better guide by considering our premises and how to reason in the light of the supervaluationist's semantics.

13.5 Supervaluationism: A Better Guide to Reasoning?

According to the supervaluationist theory of vagueness, a vague sentence is true (false) iff it is true (false) on all ways of making it precise (on all 'precisifications').¹⁷ If

¹⁷ Supervaluationism is the theory of vagueness I defend in Keefe, 2000. The locus classicus of the theory is Fine, 1975.

Tek is borderline tall, 'Tek is tall' will be neither true nor false, because if we were to make 'tall' precise, we could draw a boundary at a height above Tek's or at one below his height, so it is neither true that he is tall on all ways of making 'tall' precise nor false on all those ways. The main premise of the Sorites paradox, (C_g), comes out false because on each way of making 'tall' precise, there is a sharp boundary, so a consecutive pair in the series where one counts as 'tall' and the next one doesn't. But since it is a different pair on different precisifications, there is no pair of which it is true to say the first is tall and the second isn't. The parallel premise of an unproblematic reversed argument, (R_g) 'if y is one hundredth of an inch *taller* than x , and x is tall, then y is tall' does come out true, however, as it is true however we make 'tall' precise.

To use this theory to guide our reasoning, we should consider our premises by asking whether they are true on all ways of making them precise and only go on to reason with them if so. We can then use classical logic in our reasoning as this is the logical system delivered by the supervaluationist theory of vagueness.¹⁸ Appropriately, this story about how to reason helps us assess our premises, in addition to telling us what arguments are valid. We will do well, in general, to be cautious with our reasoning in certain cases where vagueness is present, and the supervaluationist guidance on reasoning provides a clear, plausible story about how to deal with various arguments. For example, when confronted with the Sorites argument, we are told not to go ahead and reason from the premises, because (C_g) is false on all precisifications. On the other hand, though still a vague sentence, the reversed premise (R_g) can be used as a premise of an argument since it is true on all precisifications.

Since supervaluationists declare the generalized Sorites premise (C_g) false, they need to explain away our intuition that it is true. This is typically done by, for example, pointing out that although it is false, it does not have a false instance, and by offering statements (e.g. ones containing the definitely operator) that do come out true according to supervaluationism which are suitably similar to the premise as to render it plausible that there is some confusion between them, and yet which do not yield paradox.¹⁹ It is true that degree theorists can and sometimes do endorse similar such explanations (see, e.g., Edgington, 1996, p. 311). That is no point in favour of the degree theorist's framework though, which is redundant in that explanation. There may appear to be a tension between my acceptance of the falsity of the generalised premise here, on my theory, and my earlier objection to the way that that premise came out with a low degree of truth on various degree theories. But my point is that the falsity of that premise on the degree theories undermined the relatively appealing thought that vague sentences seem true when they are merely nearly true, which was at the root of the motivation for the degree-theoretic framework and the response to the Scare Story, as considered above.

Could we get the best of all worlds by introducing a degree-theoretic structure within the supervaluationist framework? Such a structure can be provided by

¹⁸ See Keefe, 2000, pp. 174–81. There are some complications when it comes to arguments involving the 'definitely' operator, but I put aside those issues for now.

¹⁹ See e.g. Keefe, 2000, pp. 184–6.

assigning a degree to a sentence according to the proportion of precisifications on which it is true.²⁰ This structure will then allow us to capture a gradual decline in degree of 'x is tall' through the Sorites series, with the early borderline cases counting as tall on most, but not all, precisifications and the late borderline cases counting as not tall on most precisifications. As with other degree-theoretic pictures, though, the degrees thereby assigned to complex sentences will not typically reflect the extent to which we are inclined to believe them (for example, the main Sorites premise will, of course, come out true to degree 0). These degrees will not help with guiding our reasoning or with answering Edgington's Scare Story. When I come to reason from some premises, I cannot, for example, trust a vague sentence that I am more inclined to believe more than one than I find less compelling as the former may be one of those many complex cases which seems true without being nearly true. To see to what extent I should trust it, I need to look to the framework of precisifications. But then, I can follow the guidance that I don't trust that premise unless it is true on all precisifications. The degrees play no role there in cases like the Sorites with its false premise or the non-paradoxical parallel argument involving the true premise (R_g).

If a premise is false on all precisifications, I should not trust it and if it is true on all, I should. What if a premise is true on some and false on others, so is neither true nor false? Could supervaluationist's degree play a role here? Consider, for example, the argument that extracts one step from the long Sorites, involving a nearly true predication of F to x_j and the individual conditional (C_j) deriving the conclusion that the next member of the series, x_{j+1} , is F . Perhaps the addition of degrees to the supervaluationist framework appropriately allows us to trust the conclusion of this reasoning to just a slight degree less than the conclusion. But, to see whether this is a significant addition to the general supervaluationist picture, we need to say more about what the supervaluationist who has not introduced degrees should say about reasoning with premises that are neither true nor false.

I maintain that advice on how to evaluate reasoning with premises that are neither true nor false is a complex affair. Clearly we should be cautious of accepting the conclusion. If there is no precisification on which the premises are all true together (so the conjunction of the premises is false), then those premises provide no grounds at all for accepting the conclusion. This is the situation with the long Sorites with the many conditional premises and in this case (if not always in others) we can see this incompatibility when considering the premises together. If the premises are all true together on some precisifications and not on others, the advice is less clear-cut. But, I suggest, it is both less clear what our supervaluationist story should say about these cases and what we should *want* a good story to say about them. The latter may vary depending on the kind of situation in which an argument is offered (even when one of its premises is known to be neither true nor false). Sometimes such an argument is advanced because we want the premise to be treated *as* true—a kind of stipulation. For example, if I offer the premise ' x_m is tall' when we would normally call x_m borderline tall, that might be because I want this to be treated as true in the context.

²⁰ For discussion of this idea, see Kamp, 1975; Lewis, 1970; and Keefe, 2000, pp. 171–2. Edgington, 1996, pp. 315–16 also suggests degrees treated in this way as a heuristic device in understanding 'verities', noting that the relevant proportions of precisifications have the same probabilistic structure.

Now, we can make sense of this kind of situation in the supervaluationist framework by restricting the relevant precisifications in the context. We should then trust the conclusion in that context (i.e. given the stipulation) if the other premises are true on all those precisifications. For example, if I start from ' x_m is tall' and use (R_g) ('if y is one hundredth of an inch taller than x , and x is tall, then y is tall' to work up the series), my conclusion that the later, taller, members of the series are tall should be accepted in the context. If, on the other hand, other premises are neither true nor false (true in some of the relevant precisifications, false in others), we will not be similarly entitled to draw the conclusion. For example, if I continue down the series using (C_g) instead, then the reasoning is questionable.

So, if faced with an argument with a premise that is neither true nor false, we need to decide whether or not to treat that premise as stipulated as true, and that will require us to look to the context. Care will need to be taken, especially as with certain complex sentences we can't consistently stipulate that they are true even if we want to—the Sorites premise might be an example of this phenomenon.

This only gives the beginnings of a story to guide our reasoning and the evaluation of reasoning in the presence of vagueness, even if supervaluationism is accepted. But, again, I take this to be partly a reflection on pragmatic complexities. This does not rule out the possibility that in occasional cases, the consideration of degrees—as captured by proportions of precisifications—could have a role in guiding our reasoning or the evaluation of other reasoning. But that would not undermine the merits of the supervaluationist story, and, indeed, might be thought to build on it. Moreover, for the purposes of this paper, it should be noted that reasoning within a Sorites argument is not the kind of case where considering degrees is of any help and so we do not have a reason to pursue the analogy between the Sorites Paradox and the Preface Paradox.

13.6 Conclusion

In our reasoning, we need to be alert to both uncertainty and vagueness: the Preface and Sorites paradoxes bring out problems relating to these phenomena. Modelling our uncertainty in terms of degrees of belief gives a picture that illuminates the Preface Paradox, and can guide us in assessing different arguments where uncertainty is present. But, I have argued, a degree-theoretic account of vagueness and the Sorites paradox cannot succeed in the same way. It may initially seem to promise a similar account whereby we are told what caution to exhibit in our reasoning with vague but compelling premises (as with compelling but uncertain premises) because we can assume they should be assigned a high but non-maximal degree. But, in fact, it is obliged to say that the things that seem true or close to true are often actually not to be assigned a high degree (and in many cases are completely false). The Scare Story—that uncertainty/vagueness may render us inferentially impotent—has not been avoided just by allowing some *simple* predications to seem true while being merely nearly true if the *compound* premises we reason with do not fit this pattern.

After exploring a range of degree-theoretic approaches here, I suggested abandoning the analogy between the two paradoxes and the degree-theoretic modelling that helps in one case. I then briefly argued that in dealing with vagueness, we should

employ the supervaluationist theory in assessing both our premises and our forms of reasoning. There is no need to introduce degrees of truth or Edgington's verities and without them we have a better guide to reasoning.²¹

References

- Adams, E.W. (1966) Probability and the logic of conditionals. In J. Hintikka and P. Suppes (eds), *Aspects of Inductive Logic*. Amsterdam: North Holland.
- Cook, R. (2002) Vagueness and mathematical precision, *Mind* 111: 226–47.
- Edgington, D. (1992) Validity, uncertainty and vagueness, *Analysis* 52: 193–204.
- Edgington, D. (1996) Vagueness by degrees, in R. Keefe and P. Smith (eds), *Vagueness: A Reader*, Cambridge, MA: MIT Press, pp. 294–316.
- Field, H. (2009) What is the normative role of logic? *Proceedings of the Aristotelian Society*, Supplementary Volume 83: 251–68.
- Fine, K. (1975) Vagueness, truth and logic, *Synthese* 30: 265–300.
- Foley, R. (1992) The epistemology of beliefs and the epistemology of degrees of belief, *American Philosophy Quarterly* 29: 111–24.
- Forbes, G. (1983) Thisness and vagueness, *Synthese* 54: 235–59.
- Hawthorne, J. and Bovens, L. (1999) The preface, the lottery and the logic of belief, *Mind* 108: 241–64.
- Kamp, J. A. W. (1975) Two theories about adjectives. In E. L. Keenan (ed.), *Formal Semantics of Natural Language*. Cambridge: Cambridge University Press.
- Kamp, J. A. W. (1981) The paradox of the heap'. In U. Mönnich (ed.), *Aspects of Philosophical Logic*. Dordrecht: Reidel.
- Keefe, R. (2000) *Theories of Vagueness*. Cambridge: Cambridge University Press.
- Keefe, R. (2012) Modelling vagueness: What can we ignore?, *Philosophical Studies* 161: 453–70.
- Lewis, D. (1970) General semantics, *Synthese* 22: 18–67.
- Machina, K. (1976) Truth, belief and vagueness, *Journal of Philosophical Logic* 5: 47–78. Reprinted in R. Keefe and P. Smith (eds), *Vagueness: A Reader*, Cambridge: MA: MIT Press.
- Makinson, D. C. (1965) The paradox of the preface, *Analysis* 25: 205–7.
- Peacocke, C. (1981) Are vague predicates incoherent? *Synthese* 46: 121–41.
- Smith, N. J. J. (2008) *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.
- Williamson, T. (1994) *Vagueness*. London: Routledge.

²¹ A version of this chapter was presented to the 'Probability and Vagueness' conference at the University of Tokyo: many thanks to the participants there for helpful questions and comments. I am also very grateful to Nick Jones and Lee Walters for detailed comments on earlier drafts of this chapter.

Hysteresis Hypotheses

Alan Hájek

Edgington¹ (1992) argues forcefully that ‘the Sorites Paradox and the Lottery Paradox are *structurally* the same phenomenon. Each is a pathological case of a phenomenon we need to understand: reasoning with uncertainty, and reasoning with vague concepts’ (p. 202). Moreover, as she observes (e.g. 2011), there are important parallels between the Lottery Paradox and the Preface Paradox. In this chapter I assimilate the Sorites Paradox and the Preface Paradox, drawing further parallels between reasoning with uncertainty and reasoning with vague concepts. More precisely, I discuss experiments in which subjects are taken along soritical series of coloured patches, displaying so-called *reverse hysteresis* in their responses. (I shall explain this shortly.) I offer an explanation of why reverse hysteresis is rational there. I present a variant of the Preface Paradox—the Progressive Preface Paradox—which I argue is analogous to the Sorites Paradox and its associated experiments, and offer an analogous explanation of why reverse hysteresis is again rational. The explanation’s central idea is that ‘belief’ is context-dependent.

14.1 A Sorites Experiment

Before you is a series of coloured patches, beginning on the left with patches that are pure blue, ever-so-gradually becoming greener as you move to the right, and ending with pure green patches. Adjacent patches are so similar in colour that you have difficulty discriminating between them. We take you on a ‘forced march’, asking you to label the patches successively. Your task starts easily: ‘blue, blue, blue...’, you say confidently, as we point to the early patches in the series in turn. Then your task gradually becomes harder, your confidence diminishing accordingly. You start entering the grey area, as it were. For a while you continue to call the patches ‘blue’, though increasingly uneasily. Then at a certain point you switch: ‘green, green, green...’ you say, working your way ever more confidently through the green zone until the end.

¹ It is an honour and a pleasure to contribute to this volume dedicated to Dorothy. I have admired her work for many years. I especially got to know and admire *her* during the 2009 Summer School on Conditionals at Central European University, Budapest, a high-water mark in my intellectual life. I thank Sharon Berry, Rachael Briggs, Kenny Easwaran, Paul Égré, Frank Jackson, Hanti Lin, Daniel Nolan, Diana Raffman, Wolfgang Schwarz, and Lee Walters for very helpful discussion and comments on this chapter.

Now, we do things in reverse. We start you at the green end and ask you to label the patches from right to left. ‘Green, green, green, ...’, you confidently begin your successive labelling. But as you proceed further, your judgments become more tentative. Nevertheless, you persist for a while: ‘green, green, green, ...’ Then at a certain point you switch: ‘blue, blue, blue ...’, growing in confidence until you reach the left-most patch.

The key thing we notice is that there are some patches (around the middle) that you label as ‘green’ when you approach them from the left, and as ‘blue’ when you approach them from the right. You switch labelling closer to where you started: the point at which you switch is further to the left as you are marched rightwards than it is when you are marched leftwards. You display *reverse* (or *negative*) *hysteresis* in your categorisations. (It is so-called by contrast to *hysteresis*, in which one *persists* longer with the categorisation with which one started.) Moreover, we may suppose that you are authoritative about your perceptual states, and that you are being sincere at every point. It’s as if you are anticipating the colour that is waiting for you at the other end. Or it’s as if you readily revise your initial judgments.

You are in good company: Kalmus (1979), Égré, de Gardelle, and Ripley (2013), and Raffman (MS and 2014) all have found that experimental subjects taken on similar forced marches tend to display reverse hysteresis, just like you. Moreover, your company really does seem to be *good*: this pattern of behaviour is entirely rational. Soon I will suggest why.

14.2 The Progressive Preface Paradox

Imagine the following situation. You have just written a long book. You painstakingly researched each sentence in it; you wouldn’t have committed it to the page otherwise. Yet you know that it would be remarkable if your book were completely error-free. So you write a modest preface: ‘While I believe each sentence in the book, there is surely at least one error somewhere or other in the book.’ Let us suppose that your book consists of the sentences $S_1, S_2, \dots, S_n$. You believe S_1 ; you believe S_2 ; ... you believe S_n . Yet you do *not* believe the conjunction of all these sentences—indeed, you believe that $S_1 \& S_2 \& \dots \& S_n$ is false. The preface paradox arises from the apparently innocuous assumption that all of these beliefs may be rational. How can it be rational to hold jointly inconsistent beliefs? In particular, it seems that rational belief is closed under conjunction: if you believe p and you believe q , then you believe p and q , on pain of irrationality. But then rationality requires you to believe the conjunction of all the sentences in your book, which you do not.

We may go further. We take you on the following ‘forced march’ through your book. We ask you if you believe S_1 . ‘Yes, of course’, you reply. ‘I said that I believe each sentence in the book.’ We ask you if you believe $S_1 \& S_2$. Again you say ‘yes’. We ask you if you believe $S_1 \& S_2 \& S_3$. Again you say ‘yes’. And so we go, progressively conjoining the next sentence, and then the next ...—we may imagine this sequence of ever-lengthening conjunctions being laid out from left to right. You keep saying ‘yes’ for a while, but increasingly uneasily. Eventually, you switch: we finally confront you with a sufficiently long conjunction that you profess *not* to believe. And so it goes from then on: as we continue to add further conjuncts, you continue to profess not to

believe the resulting conjunction. In the end we reach the conjunction of all the sentences in your book, $S_1 \& S_2 \& \dots \& S_n$, which of course you modestly claimed not to believe (indeed, to disbelieve).

Now, we do things in reverse. We start by asking whether you believe $S_1 \& S_2 \& \dots \& S_n$. Of course you say 'no'. We *remove* the final conjunct, asking whether you believe $S_1 \& S_2 \& \dots \& S_{n-1}$. Again you say 'no'. And so we go, progressively removing a conjunct, and then another one, from the end. You keep saying 'no' for a while, but increasingly uneasily. Eventually, you switch: we finally confront you with a sufficiently *short* conjunction that you profess *to believe*. And so it goes from then on: as we continue to subtract further conjuncts, you continue to profess to believe the resulting conjunction. In the end we reach the first sentence in your book, S_1 , which of course you claim to believe.

The key thing we notice (in our story) is that there are some 'middling length' conjunctions that you claimed *not to believe* when you approached them from the left, with ever-lengthening conjunctions, and that you claimed *to believe* when you approached them from the right, with ever-shortening conjunctions. Moreover, your claims reflect your genuine judgments: you are authoritative about your doxastic states, and you are being sincere at every point. Your judgments display *reverse hysteresis*. The point at which you switch is further to the left as you are marched rightwards than it is when you are marched leftwards. It's as if you are anticipating the judgment that is waiting for you at the other end. Or it's as if you readily revise your initial judgments.

I submit that such reverse hysteresis may be entirely rational. Unlike the sorites series, I am not aware of any experiment that bears on my variant of the preface paradox. Such an experiment would be somewhat unwieldy to perform, and I have no intention to perform it. In any case, I don't need to perform it. I can make the case for the *rationality* of such reverse hysteresis without such empirical findings.² And soon I will.

Yet it seems to have an even more shocking consequence than the original preface paradox: it is rational for you to have inconsistent belief reports with respect to *the very same conjunction* (for some conjunction of middling length). It was bad enough that in the original paradox, you had inconsistent belief reports, but at least they were spread over a large number of sentences. How can it be rational to have inconsistent belief reports with respect to a single *sentence*?! Let me call this the *progressive preface paradox*.

The assumption that you are authoritative and sincere about your doxastic states seems innocuous enough³ (idealising suitably), but it leads to a yet more startling consequence: you believe and you don't believe the very same sentence!

² Indeed, I doubt that the original preface paradox has much *empirical* support. We are supposed to *imagine* an author writing a modest preface, and we intuit that it is *rational* for her to do so—never mind whether or not people actually write such prefaces, or are committed to them.

³ Epistemicists about vagueness such as Williamson (1994) may not find the assumption of authority so innocuous, especially when coupled with a Williamsonian anti-luminosity thesis regarding mental states. As I will argue, 'belief' is vague. Epistemicists will insist that there is always a sharp boundary between what one believes and what one does not believe, with any borderline cases being due to one's ignorance of which doxastic state one is really in. Thanks here to Lee Walters and Wolfgang Schwarz.

14.3 The Lockean Thesis

The *Lockean Thesis* is a putative bridge between binary *belief*, and *degrees of belief*. Roughly, it says that belief is sufficiently high degree of belief. A simple precisification of this idea is:

(Simple Lockean Thesis)

For all S , for all p ,

S believes that p iff S 's credence in p is at least t .

t must be set to some specific value—say, 0.95—in order for the biconditional to be at all plausible. The Thesis cannot merely say that there is *some* threshold t such that S 's credence in p is at least that. Of course there is: 0 is one such threshold; -17.358 is another. Moreover, the threshold needs to be *high*; and according to the Simple Lockean Thesis, it is sharp. Let us suppose, then, that t is 0.95; but if you prefer some other value, you can replace what I say with that value without making any substantive difference to my points.

On this view, a rational agent should not display reverse hysteresis when taken on marches forwards and backwards through her book, the way we imagined you doing above. Let us suppose that her credence in each sentence taken individually exceeds 0.95, that her credence in the conjunction of all the sentences is below 0.95, and that this is the appropriate threshold in the Simple Lockean Thesis.⁴ Then when we start with the book's opening sentence, and successively conjoin more and more sentences to it, there must be a first point at which her credence in the resulting conjunction falls below 0.95. That is the exact point at which the switch from belief to non-belief must occur. When we go through this process in reverse, the switch point is the same. (I assume that your credences in the sentences and in their conjunctions have not changed in the interim.) So much for the putative rationality of the reverse hysteresis, you might say.

Moreover, the Simple Lockean Thesis affords the simplest of solutions to the original preface paradox. You believe each sentence taken individually—your credence in it clears the threshold. You believe the conjunction of all the sentences is false—your credence in the negation of the conjunction clears the threshold. A long conjunction of very probable sentences can easily be very improbable, and then it has a very probable negation. Rational belief is not closed under conjunction—period. So much for the preface 'paradox', you might say.

So much for the Simple Lockean Thesis, we should say. It is *too* simple to be credible. Indeed, ironically, the very simplicity of its 'solution' to the preface paradox arguably backfires, indirectly providing an argument *against* the Thesis! For the intuitions that give the paradox its force, and in particular in favour of the closure of rational belief under conjunction, are not so easily dismissed. Whatever our error may be when we are gripped by the paradox, it is surely subtle. A good solution to the paradox should expose the error, but also illuminate why it is tempting—why we are

⁴ Let us also suppose that all of her credences are sharp, so that we don't have to fuss with what the Lockean Thesis looks like when credences are imprecise.

so easily disposed to make it.⁵ To be sure, high probability is not closed under conjunction; but that is hardly a subtle point. It should be obvious after a moment's reflection, in a way that the solution to the preface paradox is not.

Moreover, I think it is implausible that the line between belief and non-belief is as sharp as the Simple Lockean Thesis has it.⁶ A lottery with tickets numbered 1 through 100 is about to be played. According to the Thesis, you *believe* that the winning ticket will have a number from 1 through 95, but you *don't believe* that it will have a number from 1 through 94. Frankly, I find this hard to believe.

I regard another strike against the Thesis to be the all-too-decisive way in which it rules out the rationality of reverse hysteresis in the progressive preface paradox. However, I need to make my case *for* this rationality. Before I do that, I need to make my case for such reverse hysteresis being rational in the sorites experiment. And I will do *that* after briefly considering some alternative views.

14.4 The Rationality of Reverse Hysteresis in Our Original Sorites Experiment

You judge that a certain patch is blue when you approach it one way, and green when you approach it the other way, reverse hysteretically. How do we explain your giving these opposed judgments? And how do we explain your judgments displaying reverse hysteresis? Better still, can we *rationalise* your judgments, rather than merely explaining them?

Offhand, we might expect the exact opposite pattern of judgments: *hysteresis*. After all, it seems this pattern would be easy to rationalise. Having just labelled a disc a certain colour, when moving to a neighbour of barely distinguishable colour, you should feel pressure to give it the same label. So when you start at the blue end, it seems you should persist in labelling the discs 'blue', only switching when you really feel forced to; and similarly, when starting at the green end, you should only switch from 'green' when you really feel forced to. And yet we are imagining that you, like the experimental subjects, don't do this. It is reverse hysteresis, rather than hysteresis, that we want to rationalise. It seems that we have our work cut out for us!

There would be no problem if the 'you' who judges the patch to be blue is in some relevant sense different from the 'you' who judges it to be green (much as there is no problem if *you* judge it to be blue and *I* judge it to be green). Of course, the time-slice of you who makes the first judgment is different from the time-slice of you who makes the second judgment.⁷ But we surely do not want to individuate epistemic agents so finely. *You* persist from one experimental condition to the next, even if your judgment regarding a particular patch does not.

⁵ Discussion with Wolfgang Schwarz helped me here.

⁶ An epistemicist about vagueness might find this plausible, insisting that this line *is* sharp, even though we do not know where it is.

⁷ If you think that judgments take some time to make, and that time slices are not sufficiently temporally extended to make them, I can make the point in terms of time *slivers* of you, which *are* sufficiently extended, though thin enough to differ from one judgment to the next.

Perhaps, however, the way the patch appears to you changes from its first to its second appearance. Égré's (2011) account of reverse hysteresis in the sorites experiments is that 'borderline cases are ambiguous items between two adjacent categories: they can legitimately be viewed in two distinct and even opposite ways, though to various degrees' (65). He, in turn, is strongly influenced by Raffman (1994) (in turn influenced by Barcan Marcus) likening a category shift along a sorites series to a 'duck/rabbit' Gestalt shift. Here is a crucial passage from Raffman (in which 'blue/green' is replaced by 'red/orange'):

... one and the same *look* looked red before the shift, orange afterward; and similarly for some of the preceding patches. Ruth Barcan Marcus has suggested to me that such an effect may amount to a kind of Gestalt switch: there is a similar respect in which (for example) the duck-rabbit 'looks the same' while yet 'looking different' as its species fluctuates. I want to urge that both cases of shift mobilize a process of seeing as: just as we can see the duck-rabbit either as a duck or as a rabbit, we can see [patch] #26 and the others either as looking red or as looking orange. It is because we undergo such shifts that we perceive no discontinuities in the series. (53)

This is an ingenious proposal, but it does not meet my needs. My goal, remember, is to rationalise reverse hysteresis. Firstly, this proposal does not *rationalise* reverse hysteresis. Secondly, it does not rationalise *reverse* hysteresis. (To be sure, Égré's and Raffman's goals may be different from mine.) Let me take these points in turn.

Firstly, while it may be an adequate *psychological* explanation of what we *in fact do*, it does not explain why it is *reasonable* for us to do it. For all the explanation says, even though it is legitimate to view a borderline case in two distinct ways, having fixed on one of these ways, it may then be *illegitimate* to view it the opposite way. (Compare: while it may be legitimate for you to give high credence to *p* and for me to give low credence to *p*, given our priors, it may be *illegitimate* for you to switch from high credence to low credence, unless you get some evidence to justify doing so.) The explanation needs to go further, illuminating why *switching* judgments is legitimate (beyond the legitimacy of *having* one judgment, or the other); and better still, why such switching should even be expected of a rational agent.

Secondly, Égré's account does not explain why subjects display *reverse* hysteresis rather than hysteresis. We could equally regard hysteresis, if it were to be displayed, as involving a Gestalt shift. The question remains: why is there one distinctive pattern to the switching, rather than the opposite? So even as a psychological explanation, it is incomplete.

Here, then, is my *rational* explanation of *reverse* hysteresis. Note that it is compatible with Égré's account, though not entailed by it.

We start you with a paradigmatically blue patch, making salient and reinforcing what true blue is like. This primes you to have a demanding standard for what counts as 'blue'; this patch easily meets it. This is a prototype of something clearly blue on which you imprint—a blueprint. But as you advance to the right, the patches becoming noticeably less and less clear cases of blueness, you comparatively quickly reach patches that fail to meet the standard. Their *falling short* of the standard is salient. Suitably attentive to this, you switch categorisation comparatively early. Then we start you at the other end, priming you to have a demanding standard for

what counts as 'green'. A prototype of something clearly green is loaded up in your mind. But as you advance to the left, the patches becoming noticeably less and less clear cases of greenness, you comparatively quickly reach patches that don't meet the standard. Suitably attentive to this, you switch categorisation comparatively early again.

Said another way: you begin your rightwards journey in a context in which the word 'blue' picks out a high-standard, blue-ribbon concept of blue—call it BLUE_{HIGH}. Comparatively few of the patches fall under the BLUE_{HIGH} concept, so comparatively quickly you rightly stop calling them 'blue'. And you begin your leftwards journey in a context in which the word 'green' picks out a high-standard concept of green—call it GREEN_{HIGH}. Comparatively few of the patches fall under the GREEN_{HIGH} concept, so comparatively quickly you rightly stop calling them 'green'.

This explanation is tailor-made for the case in which you begin your march with a paradigmatically blue patch. However, what should you do if your march begins slightly further down the series, with a slightly greenish-blue patch? If you see that you are being marched inexorably towards green patches, perhaps you will still display reverse hysteresis? This time, however, you did not imprint on BLUE_{HIGH}. What am I to say about this case?

I don't know of any experiment that tested a case like this, although as before, what matters to me is the *rationality* of a pattern of behaviour, whether or not subjects actually display it (although if they did display it, it would be nice to rationalise it). I think that a similar rationalisation of reverse hysteresis still goes through. This time you still imprint on a pretty high standard of blue—call it BLUE_{PRETTY HIGH}—which you comparatively quickly leave. Soon the patches fall short of meeting even the standard for BLUE_{PRETTY HIGH}. Accordingly, your judgment switches comparatively quickly to 'green'. And so it would go if instead we started you on a merely somewhat blue patch that is noticeably tinged with green, and marched you towards the green end. For having imprinted on *that* somewhat high standard of blue—call it BLUE_{SOMEWHAT}—you will quickly encounter patches that fall short even of that standard. And yet again, reverse hysteresis would be rational.

This suggests a more nuanced and more interesting reverse hysteresis hypothesis. Reverse hysteresis comes in degrees. If you switch categorisations at the same place when marched rightwards as you do when marched leftwards, you display no reverse hysteresis at all. If the former march's switch point is shifted just slightly to the left of the latter march's switch point, you display slight reverse hysteresis. If the switch point is considerably shifted, you display considerable reverse hysteresis. And so on. We can even *quantify* reverse hysteresis by how many patches lie between the rightwards-march switch point and the leftwards-march switch point. Now, the hypothesis:⁸ it is rational for you to display *greater* reverse hysteresis, the *higher* the standard of blueness of your first patch. If you start with a high-standard blue patch, it is rational for you to display a high degree of reverse hysteresis; if you start with a pretty high standard blue patch, it is rational for you to display a pretty high degree of

⁸ I am grateful to Hanti Lin for making it explicit.

reverse hysteresis; and so on. As the degree of blueness of your first patch goes down, so does your degree of reverse hysteresis.

But what about the countervailing pressure towards *hysteresis* that I described earlier? The fact that you can barely discriminate adjacent patches militates in favour of your persisting with the same colour label. It seems, then, that you are in the centre of a tug-of-war between two opposing forces, one pulling you towards reverse hysteresis, the other towards hysteresis. How do we rationalise your ceding to the former force rather than the latter? Arguably, the latter force is not so strong after all.⁹ For you know where the sorites series is inevitably leading you. You realise that sooner or later you will have to jump ship, and switch your colour labelling. It might as well be sooner; it is no more rational to make it later. To display hysteresis is just to delay the inevitable. This realisation weakens the force towards hysteresis. By contrast, the force towards reverse hysteresis that comes from imprinting on a pure prototype of one colour, or the other, retains its strength, and it ultimately wins.

14.5 The Rationality of Reverse Hysteresis in the Progressive Preface Paradox

Recall the heart of the problem to which reverse hysteresis gives rise in the progressive preface paradox: there is a particular conjunction—call it *p*—that you claim to believe when approaching it from the left, but that you claim not to believe when approaching it from the right. You report inconsistent attitudes to this conjunction. Moreover, we take your reports to be trustworthy. How can this be?

Belief reports have the form: *S* believes *p*. There are three moving parts here: *S* ('you'), believes, and *p*. Perhaps we can make them move a little more?

Moving 'you': There would be no problem if the 'you' who believes the conjunction *p* when it is presented in one direction is in some relevant sense different from the 'you' who does not believe it when it is presented in the other direction (much as there is no problem if *you* believe it and *I* don't). Of course, the time-slice who apprehends *p* the first time is different from the time-slice who apprehends *p* the second time, because the time has changed. Moreover, 'time-slice epistemologists' insist that 'the fundamental norms of rationality are temporally local' (Moss 2015, 172). In particular, they deny that there are diachronic norms of rationality, such that what one time-slice of yours believes is beholden to what a previous one believed. Perhaps such a wafer-thin notion of an epistemic agent will allow your judgements regarding *p* to change, depending on the direction from which you approach it?

To be sure, the two 'you's who apprehend *p* are different in some sense; but they are not different in a relevant sense. In particular, there is no relevant difference in their *evidence*. For example, we suppose that you do not lose any of your evidence by forgetting or some other cognitive mishap. (Forgetting is one of the main motivations for time-slice epistemology.) Nobody tries to solve the original preface paradox by appealing to *n* different agents who believe *n* different sentences of your book, and a

⁹ Thanks here to Wolfgang Schwarz.

further agent who disbelieves their conjunction! That would be absurd, because we suppose that there is no relevant difference in the evidence of all these agents, even if we distinguish them. And all of them are beholden to their evidence.

Moving p: We might instead appropriate Égré's account. *p* is a borderline case of a belief of yours. Perhaps it is ambiguous between belief and non-belief: you can legitimately view it as a belief, or as a non-belief. In particular, perhaps *p* appears to you as a non-belief on its first appearance (marching you rightwards), and as a belief on its second appearance (marching you leftwards). You believe it under its first guise—call it *p*₁; you don't believe it under its second guise—call it *p*₂. Perhaps—but as before, this would not *rationalise* your *reverse* hysteresis. Can we do better?

Moving believes: 'Belief' is vague. 'Believes' is a sorites-susceptible predicate, and there are apparently borderline cases of belief—two of the chief hallmarks of vagueness. The progressive preface paradox brought out both of these aspects, I hope. Your growing unease as we marched you through ever-lengthening conjunctions of sentences from your book was just like your growing unease as we marched you along the sorites series of patches. Your unease hit its peak when you were in the thick of the apparently borderline cases of belief/blue, and then abated as you moved out of them. The lottery argument against the Lockean Thesis above only brings out the vagueness of 'belief' further. Even if we allow that beliefs generally correspond to high credences, and non-beliefs to lower credences, it is implausible that a tiny change in credence could give rise to a difference in whether there is a corresponding belief or not. The fixed, sharp line that the Thesis draws between beliefs and non-beliefs is as implausible as a fixed, sharp line between heaps and non-heaps, or between blue and green patches in a sorites series.

A number of authors in the vagueness literature maintain that vague predicates are context-dependent. Kamp (1981) writes: '[i]t is typical of a vague predicate that what objects it is true of depends on the context in which it is used' (242). Raffman (1996) is even more forthright, dropping the hedge of 'typical', and making her first 'datum' about 'the correct application of a vague predicate' that it 'varies with context' (175). I think that we should keep the hedge—for arguably not *all* vague predicates are context-dependent. Consider a vague, context-dependent predicate in which all the contextual parameters are made explicit.¹⁰ The result is a vague, context-independent predicate. *Tall* is vague and context-dependent. Now, pile on every contextual parameter to which you think it is sensitive in a particular case: *tall-for-a-non-basketball-playing-third-grade-Japanese-boy*, or what have you. (I am supposing that at least in principle this can be done.) The resulting predicate is no longer context-dependent, but it is still vague: it is sorites-susceptible, and it admits of borderline cases. That said, vague predicates are *typically* context-dependent. It would be surprising if 'belief' were not. In that respect, it is surely like 'blue' or 'green'.

Suitably receptive to 'belief' being context-dependent, and suitably primed by my rationalisation of the reverse hysteresis in the sorites experiment—its appeal to priming, in particular—we are now ready for my rationalisation of it in the progressive preface paradox.

¹⁰ I am indebted to discussion with Daniel Nolan here.

We began by marching you forwards through your book, with ever-growing conjunctions of its sentences. We started you with a paradigm case of a sentence that you believed, making salient and reinforcing what pure belief (as it were) is like. This primed you to have a demanding standard for what counts as ‘belief’; this sentence easily met it. A prototype of a clear belief was loaded up in your mind. But as you advanced to the right, the conjunctions becoming noticeably less and less clear cases of things that you believed, you comparatively quickly reached conjunctions that failed to meet the standard. Their *falling short* of the standard was salient. Suitably attentive to this, you switched categorisation comparatively early. Then we started you at the other end, priming you to have a demanding standard for what counts as ‘non-belief’. A prototype of a clear non-belief was loaded up in your mind. But as you advanced to the left, the conjunctions becoming noticeably less and less clear cases of non-belief, you comparatively quickly reached conjunctions that didn’t meet the standard. Suitably attentive to this, you switched categorisation comparatively early again.

Said another way: you began your rightwards journey in a context in which the word ‘belief’ picked out a high-standard conception of belief—call it BELIEF_{HIGH}. Comparatively few of the conjunctions met the BELIEF_{HIGH} standard, so comparatively quickly you rightly stopped saying that you ‘believed’ them. And you began your leftwards journey in a context in which ‘non-belief’ picked out a high-standard conception of non-belief—call it NON-BELIEF_{HIGH}. NON-BELIEF_{HIGH} corresponds to a *low* standard for belief—call it BELIEF_{LOW}. (When belief is comparatively easy to come by, non-belief is comparatively hard to come by.) Comparatively few of the conjunctions met the NON-BELIEF_{HIGH} standard, so comparatively quickly you rightly stopped saying that you ‘did not believe’ them.

This explanation is tailor-made for the case in which you begin your rightwards march with a paradigm case of a belief. This is a reasonable assumption: in the original preface paradox, we may imagine that each sentence in your book is a paradigm case of a belief, hence the first sentence is. However, what should you do if your march begins slightly further down the series, with a conjunction of sentences that you do not believe so clearly? Much as before, let’s assume you still imprint on a pretty high standard of belief—call it BELIEF_{PRETTY HIGH}—which you comparatively quickly leave. Soon the conjunctions fall short of meeting even the standard for BELIEF_{PRETTY HIGH}. Accordingly, your judgment switches comparatively quickly to ‘do not believe’. Similarly, if we start you still further down the series, at a conjunction that you believe even less clearly but that still meets the standard for BELIEF_{SOMEWHAT}—you will quickly encounter conjunctions that fall short even of that standard. Again, reverse hysteresis would be rational.

And we may repeat the observation that reverse hysteresis comes in degrees. We can even *quantify* it by how many conjunctions lie between the rightwards-march switch point and the leftwards-march switch point. Much as before, I hypothesise that it is rational for you to display *greater* reverse hysteresis, the *higher* the standard of belief your starting point meets. If it meets a high-standard for belief, it is rational for you to display a high degree of reverse hysteresis; if it meets a lower but still pretty high standard, it is rational for you to display a pretty high degree of reverse hysteresis; and so on. As the standard goes down, so does your degree of reverse hysteresis.

Much as before, you are caught in the middle of two opposing forces. There is a force towards hysteresis that arises from your wanting to classify alike adjacent conjunctions that you can barely distinguish in credence. And there is a force towards reverse hysteresis that arises from your imprinting on a high standard for belief. As before, the former force is not so strong after all; as before, the latter force wins, and rightly so.

This also resolves the paradox that you believe and you don't believe the very same sentence assuming that you are authoritative and sincere about your doxastic states. You BELIEVE_{LOW} the sentence, but you don't BELIEVE_{HIGH} the sentence. No inconsistency there.

The context-dependence of belief that I am suggesting will have ramifications for other concepts in which belief is implicated. In particular, plausibly it will yield context-dependence of *knowledge*.¹¹ After all, plausibly belief is a necessary condition for knowledge, and typically the context-dependence of a necessary condition for something is inherited by the thing itself. Note that this is a different kind of contextualism about knowledge from contextualism based on what possibilities can reasonably be ignored (Lewis, 1996), or what relevant alternatives there are (Dretske, 2000). It is closer to DeRose's (1995) contextualism based on shifting standards for knowledge claims.

Finally, the context-dependence of belief casts yet more doubt on the Simple Lockean Thesis. It is a familiar point that a single, fixed threshold cannot respect intuitions both about beliefs in everyday propositions and lottery propositions. I have various beliefs in ordinary propositions for which my credences are noticeably below 1. I believe that Khartoum is the capital of Sudan, that it will be sunny in Los Angeles tomorrow, and that Edgington argued forcefully in 1992 that the Sorites Paradox and the Lottery Paradox are structurally the same phenomenon. But I give credence less than 0.999 to each of these propositions.¹² In particular, I give higher credence to each ticket losing in a 1001-ticket lottery. Yet I do not believe of each ticket in that lottery that it will lose. Setting the putative Lockean Thesis bar at 0.999 is both too high to respect the former beliefs, and too low to respect the latter non-beliefs. No single bar can respect all of them.

However, I think there is *something* right about the Lockean Thesis's identification of beliefs with high probabilities, and non-beliefs with lower probabilities. It seems that lotteries, whose chancy natures are so salient, set up special contexts in which the bar for belief gets set especially high. So the bar must be sensitive to the nature of proposition that is believed (or not), or the partition of propositions in which it appears.

I submit that the bar must also be sensitive to the nature of the *belief* at issue. There are variable standards for what counts as a belief, even in everyday propositions. The standards for BELIEF_{LOW} are lower than they are for BELIEF_{HIGH}, and there are doubtless various other concepts of belief with yet other standards—BELIEF_{SOMEWHAT}, BELIEF_{PRETTY HIGH}, and so on. This is not a matter of the

¹¹ Thanks here to Lee Walters.

¹² I'm sure that she argued forcefully, less sure that she did so in 1992!

contents of these attitudes—the date of an argument of Edgington’s as opposed to the fate of a ticket in a lottery. You can bear one of these attitudes and not bear another one to the very same sentence—BELIEVING_{LOW} a middling-length conjunction of sentences in your book, but not BELIEVING_{HIGH} it. No single Lockean Thesis bar can capture both attitudes simultaneously. More generally, no single Lockean Thesis bar can capture the various concepts of belief, with their various standards, simultaneously.

References

- DeRose, K. (1995) Solving the skeptical problem, *Philosophical Review* 104(1): 1–52. Reprinted in K. DeRose and T. A. Warfield (eds), *Skepticism: A Contemporary Reader*, Oxford: Oxford University Press.
- Dretske, F. I. (2000) The pragmatic dimension of knowledge. In *Perception, Knowledge and Belief: Selected Essays*, Cambridge: Cambridge University Press.
- Edgington, D. (1992) Validity, uncertainty, and vagueness, *Analysis* 52(4): 193–204.
- Edgington, D. (2011) Handout for ‘Three Grades of Uncertainty’, http://www.uniss.it/documenti/Three_grades_of_uncertainty_handout_-_Complete.pdf
- Égré, P. (2011) Perceptual ambiguity and the sorites. In R. Nouwen et al. (eds), *ViC 2009*, LNAI 6517, Berlin, Heidelberg: Springer-Verlag, pp. 64–90.
- Égré, P., de Gardelle, V., and Ripley, D. (2013) Vagueness and order effects in color categorization, *Journal of Logic, Language, and Information* 22, 391–420.
- Kalmus, H. (1979) Dependence of colour naming and monochromator setting on the direction of preceding changes in wavelength, *British Journal of Physiological Optics* 32(2), 1–9.
- Kamp, H. (1981) The paradox of the heap, in U. Mönnich (ed.), *Aspects of Philosophical Logic*, Dordrecht: D. Reidel, pp. 225–8.
- Lewis, D. (1996) Elusive knowledge, *Australasian Journal of Philosophy* 74: 549–67.
- Moss, S. (2015) Time-slice epistemology and action under indeterminacy, In T. Szabó Gendler and J. Hawthorne (eds), *Oxford Studies in Epistemology*, Oxford: Oxford University Press, pp. 172–94.
- Raffman, D. (1994) Vagueness without paradox, *Philosophical Review* 103: 41–74.
- Raffman, D. (1996) Vagueness and context-relativity, *Philosophical Studies* 81: 175–92.
- Raffman, D. (2014) *Unruly Words: A Study of Vague Language*, Oxford: Oxford University Press.
- Raffman, D. (MS) Tolerance and the competent use of vague words.
- Williamson, T. (1994) *Vagueness*. London: Routledge.

Verities and Truth-values

Nicholas K. Jones

Studying for a PhD under Dorothy Edgington's supervision was a great intellectual pleasure.* She was a patient, dedicated, rigorous, and inspiring teacher. Both in person and in print, Dorothy provides a model for me of how philosophy should be done. It is a pleasure to acknowledge my debt with this study of her thinking about vagueness.

A powerful motivation for degree-theoretic semantics is to semantically capture one of the most striking features of vagueness: the gradual transition from uncontroversial truth to uncontroversial falsity along an appropriate series. The devil, as ever, is in the details. Standard many-valued semantic theories cannot capture penumbral connections, deliver unattractive logics, and violate such seeming truisms as the T-biconditionals. Edgington's theory promises a solution to these difficulties. Edgington uses an analogy between the sorites and lottery paradoxes to motivate a probabilistic version of degree theory that builds penumbral connection in at the ground-floor, does not invalidate classical logic, and in which the T-biconditionals hold without restriction. This chapter examines Edgington's view. My primary focus will be the relationship between Edgington's degrees and truth, in light of her claim that the former do not disturb or displace classical bivalent truth.

Section 15.1 introduces Edgington's view. I describe her probabilistic treatment of the connectives in Section 15.1.1, her account of validity in Section 15.1.2, and her solution to the sorites in Section 15.1.3. The rest of the chapter concerns the relationship between Edgington's framework and classical semantics. One of the central questions is whether, and in what sense, the many degrees assigned to sentences in Edgington's framework are new truth-values to replace the classical two. Section 15.2 examines and rejects Edgington's intriguing suggestion that her framework and the classical framework provide non-competing descriptions of a single underlying semantic phenomenon. Section 15.2.1 introduces this idea. Section 15.2.2 articulates a notion of semantics that's neutral between Edgington's framework and the classical view. Section 15.2.3 then argues that Edgington's

* Thanks to Anil Gomes, John Hawthorne, Rosanna Keefe, Rory Madden, Ian Phillips, Peter Sutton, the members of the Hossack-Textor work-in-progress group, and the audience at the *Conditionals and paradox* conference in London for comments and discussion. I'm especially grateful to Dorothy Edgington for discussion of this material.

framework and the classical one cannot be non-competing descriptions of semantics understood in this sense. That leaves two options open at the start of Section 15.3: (a) classical truth and falsity are semantic notions, whereas Edgington's degrees are not; (b) Edgington's degrees are semantic notions, whereas classical truth and falsity are not. In recent work, Edgington endorses a primitivist view about vagueness that fits well with option (a). She argues for this view by arguing that bivalence holds without restriction, even in borderline cases, which appears to tell against option (b). Section 15.3.1 argues that Edgington's arguments for bivalence presuppose a substantive thesis about truth. Section 15.3.2 closes the chapter by arguing that this thesis is incompatible with option (a). I conclude by suggesting that the best version of Edgington's proposal will be a version of option (b).

A note on terminology before I begin. Clarity will be my theoretically neutral concept for distinguishing borderline cases from the rest. It's clear that p iff p and it isn't borderline whether p ; it's clear that not- p iff not- p and it isn't borderline whether p ; it's borderline whether p iff it's neither clear that p nor clear that not- p . I'll say that x is clearly F when it's clear that x is F . So, for example, scarlet is clearly a shade of red, pure orange clearly isn't a shade of red, and somewhere between them in the spectrum lie shades that are only borderline shades of red (and borderline shades of orange too).

15.1 Edgington's Proposal

This section introduces Edgington's view: Section 15.1.1 outlines her account of the connectives; Section 15.1.2 turns to validity; Section 15.1.3 looks at the sorites. My primary source is 'Vagueness by degrees'.¹

15.1.1 The framework

Edgington's view has four key features. Firstly, her approach is degree-theoretic: sentences are assigned degrees, represented by real numbers in $[0,1]$. Edgington calls these degrees *verities*. Throughout, the number assigned to a sentence A will be $v(A)$.² Secondly, verities are not degrees of truth, or new truth-values in competition with the classical two; verities are degrees of *closeness to clear truth*.³ Sections 15.2–15.3 examine the relation between verity and truth-value in more detail. Thirdly, greater numbers represent closer proximity to clear truth, with 1 as clear truth and 0 as clear falsity. Fourthly, the connectives and verities interact via the probabilistic rules:

$$\begin{aligned} v(\neg A) &= 1 - v(A). \\ v(A \vee B) &= v(A) + v(B) - v(A \wedge B). \\ v(A \wedge B) &= v(A) \times v(B | A). \end{aligned}$$

Here, $v(B | A)$ is the conditional verity of B given A :

¹ Edgington (1997), henceforth 'VbD'.

² I use ' A ', ' B ', ' C ' as metalinguistic variables ranging over sentences, and ' p ', ' q ', ' r ' as object-language propositional constants (though I'll also sometimes use them as placeholders for sentences in the body-text).

³ VbD, p. 299; Edgington, MS, §§2, 4; Edgington, 2010, pp. 105–6.

Conditional verity may be explained thus. Borderline cases of a vague term, e.g. 'red', with verities between 1 and 0, are cases which it would not be definitely wrong to count as red. We can hypothetically decide to count such a case as above the borderline—as definitely red—and see what consequences this hypothetical decision has for other propositions. $v(B \text{ given } A)$, the conditional verity of B given A , is the value assigned to B on the hypothetical decision to count A as definitely true.⁴

One important feature of the way we use vague language is a kind of classificatory discretion about borderline cases: one may choose to count borderline cases either way, as cases or as non-cases, provided one makes it clear to one's conversational partners that that's what one is doing and leaves oneself open to future revision as the conversation evolves.⁵ Edgington's rules for the connectives and gloss on conditional verity place classificatory discretion at the heart of her semantics.

Conditional verity induces failures of verity-functionality: the verity of a complex sentence is not a function of the verities of its components.⁶ Verity-functionality fails because $v(p) = v(q)$ is compatible with both $v(p | r) \neq v(q | r)$ and $v(r | p) \neq v(r | q)$; so substitution of ' p ' for ' q ' within a complex sentence can preserve the verities of its components without preserving the conditional verities on which the verity of the whole complex depends. An example. Suppose a and b are both borderline red, though a is only just slightly redder than b , so that:

$$\begin{aligned} v(Ra) &= v(\neg Ra) = 0.5. \\ v(Rb | Ra) &= 0.9. \end{aligned}$$

(b counts as almost clearly red, given the hypothetical decision to count something, a , just redder than it as clearly red.)

$$v(Rb | \neg Ra) = 0.$$

(b counts as clearly not red, given the hypothetical decision to count something, a , just redder than it as clearly not red.)

Hence:

$$\begin{aligned} v(Ra \wedge Rb) &= v(Ra) \times v(Rb | Ra) = 0.5 \times 0.9 = 0.45. \\ v(\neg Ra \wedge Rb) &= v(\neg Ra) \times v(Rb | \neg Ra) = 0.5 \times 0 = 0. \end{aligned}$$

These conjunctions have different verities, though they differ only by verity-preserving substitution of conjuncts. That is a good thing. Since a is redder than b , it is natural to regard one who said that a is not red and b is red as clearly mistaken, unlike one who said that both are red. Edgington's account of conditional verity thus allows her to accommodate the conceptual relations between the borderline regions of vague predicates, known as penumbral connections:⁷ hypothetical classificatory decisions about borderline cases have implications for how other borderline cases are classified under the relevant hypotheses.

⁴ *VbD*, p. 306.

⁵ Shapiro, 2006 emphasizes this phenomenon.

⁶ This differentiates Edgington from the more traditional degree theories criticized in Williamson, 1994, ch. 4.

⁷ Fine, 1975, pp. 123–5.

It is not entirely clear how the classificatory decisions Edgington uses to explicate conditional verity should be understood. The quote above suggests that they are decisions to count a sentence as clearly true, or to count a predicate as clearly applying to an object. That is also required by her solution to the sorites, which I'll describe in Section 15.1.3. Central to that solution is an account of the verities of certain material conditionals, derived from Edgington's rules for the connectives. The derivation relies on the following procedure for calculating $v(Rb|Ra)$ when b is further from clearly red than a :⁸

First, assign maximum verity to ' Ra '. Then rescale the degrees to which things further from clearly red than a count as clearly red, holding the clear non-red cases fixed. The conditional verity $v(Rb|Ra)$ is $v(Rb)$ following rescaling.

Maximum verity is clear truth. So the first step of the procedure makes ' Ra ' clearly true. Edgington wants to explicate conditional verity using classificatory decisions. So such decisions must be capable of making ' Ra ' clearly true, when it's borderline before the decision. This raises some unanswered questions for Edgington.⁹

Suppose ' Ra ' is borderline. Suppose I decide to count it as clearly true. What happens? There are two options: (a) ' Ra ' remains borderline; (b) ' Ra ' becomes clearly true. Option (a) undermines the above procedure for calculating $v(Rb|Ra)$. So some classificatory decisions cannot be of type (a).

According to option (b), my decision changed ' Ra ' from borderline to clearly true. So borderline status for sentences must be relative to time, or to the classificatory decisions in force at a conversational context.¹⁰ Now, Edgington endorses unrestricted bivalence, even for borderline cases. (Sections 15.2.1 and 15.3 discuss bivalence further.) So before my decision: ' Ra ' was either true or false. Suppose it was true. Then my decision changed ' Ra ' from borderline to clearly true, without affecting its truth-value. How did it do so? My decision presumably affected ' R 's meaning in some other way. We need to know more; a story is owed of exactly how this works.

One might attempt to avoid this commitment by maintaining that ' Ra ' was false before my decision, and likewise for any other borderline sentence. The attempt will fail. Suppose there's another vague predicate ' O ' whose extension (within the relevant domain D) comprises exactly the anti-extension (within D) of ' R '. Then if ' Ra ' is false before my decision, ' Oa ' is true. The previous paragraph's argument can now be used to show that classificatory decisions can change ' Oa ' from borderline to clearly true without affecting its truth-value. Plausibly, 'red' and 'orange' are like this: their semantic values are related to ensure that they partition a certain region of colour space. An account of how classificatory decisions affect borderline status without affecting truth-value is still required.

⁸ Edgington employs this procedure in the central paragraph of *VbD*, p. 308.

⁹ For a different criticism see Keefe, 2000, pp. 98–100.

¹⁰ Shapiro, 2006, pp. 31–6, 73–5 develops this kind of view, using extensional variation to analyse variation in clarity. Because Shapiro's extensions and anti-extensions are not exhaustive, however, Edgington cannot adopt his approach. Note also that Shapiro distinguishes context-relative *internal* clarity from context-invariant *external* clarity (though he uses 'determinacy' rather than 'clarity').

What about if ‘*Ra*’ was false before my decision? Since clear truth implies truth, my decision changes ‘*Ra*’ from false to true. So my decision changed the truth-conditional content of ‘*Ra*’ in some way. How? Again, a story is owed.

We’ve seen that Edgington is committed to three ideas. Firstly, classificatory decisions can change a sentence from borderline to clearly true. Secondly, some such decisions don’t affect the sentence’s truth-value. Thirdly, other such decisions change the sentence’s truth-conditional content. I am not claiming that these ideas are problematic. I am claiming that a central component of Edgington’s view remains under-described until an account of how this all works is supplied. Rather than getting bogged down in these issues, let us move on.

Edgington supplements the standard connectives with an operator *D*, intended to formalize ‘It is clear that...’. Since verities are degrees of closeness to clear truth and anything with verity less than 1 is at some distance from clear truth, the rule for *D* is:

$$v(DA) = 1 \quad \text{iff} \quad v(A) = 1; \text{ otherwise } v(DA) = 0.$$

Using *D*, we can say in the object-language that it’s borderline whether *p*:

$$\neg Dp \wedge \neg D\neg p.$$

Without *D*, the logic and semantics of borderline status can be studied in Edgington’s framework. With *D* in place, the logic and semantics of claims about borderline status can also be examined.

15.1.2 Validity

This section presents Edgington’s approach to validity.

Edgington introduces the *verity constraining property* (VCP) thus:¹¹

The argument from $A_1, \dots, A_n$ to *C* has the VCP iff it is not possible that: $[1 - v(C)] > [(1 - v(A_1)) + \dots + (1 - v(A_n))]$.

Call $1 - v(A)$ the *unverity* of *A*. In this terminology, an argument has the VCP iff the unverity of the conclusion cannot exceed the sum of the unverities of the premisses; that is, iff there cannot be an increase in total unverity between premisses and conclusion. Edgington argues that exactly the valid arguments possess the VCP:

Valid $\equiv$ VCP An argument is valid iff it possesses the VCP.

On Edgington’s view, the valid arguments are those in which total unverity cannot increase between premisses and conclusion. In order to understand this, we need to know what ‘possible’ means in the definition of the VCP. Edgington does not consider this directly, though her remarks about validity do suggest an answer.

Edgington uses modality to introduce validity:

By a valid argument ... I mean an argument such that it is impossible that all its premisses are true and its conclusion false.¹²

¹¹ *VbD*, pp. 302, 307.

¹² *VbD*, pp. 299–300.

Edgington does not explicate the relevant variety of possibility further. We may, however, reasonably assume that it's the same notion as she uses to characterize the VCP. On this view $\text{Valid} \equiv \text{VCP}$ says that, for an unspecified background notion of necessity, necessary possession of one of the following properties is equivalent to necessary possession of the other: (a) truth-preservation from premisses to conclusion; (b) non-increase in total unverity between premisses and conclusion. A question remains: equivalent in what sense?

It is tempting to interpret Edgington as offering a verity-theoretic analysis of validity. Temptation should be resisted. Since $\text{Valid} \equiv \text{VCP}$ follows trivially from the claim that its right-hand-side analyses its left, that interpretation renders it mysterious why Edgington offers a *proof* of $\text{Valid} \equiv \text{VCP}$.¹³ One might respond that the proof shows that the VCP is coextensive with a particular notion of validity, e.g., classical validity. But the proof requires only minimal assumptions about the formal characteristics of validity, never mind distinctively classical characteristics.¹⁴ So it cannot show any such thing.

A better interpretation takes $\text{Valid} \equiv \text{VCP}$ to capture a necessary coextensiveness between validity and the VCP, without attributing conceptual priority to one or the other. On this view, Edgington's proof of $\text{Valid} \equiv \text{VCP}$ is intended to establish that this relationship holds between the verity-theoretic properties of arguments and validity, without commitment to an explication of one in terms of the other. This allows the logical properties of arguments to be studied via their verity-theoretic properties. Edgington uses that approach to resolve the sorites. Now, although Edgington left the formal properties of modal validity unspecified, it is plausibly identified with classical validity; for that is what follows from the orthodox conception of possible worlds as maximal and consistent.¹⁵ Edgington thus offers a verity-theoretic solution to the sorites that, unlike typical degree-theoretic approaches, brings no revision in classical logic. It is to this I now turn.

15.1.3 *The sorites*

How does $\text{Valid} \equiv \text{VCP}$ help with the sorites? For brevity, I consider only the long-form sorites here.¹⁶ Let $a_1, \dots, a_n$ be a series of coloured patches with a_1 clearly red, a_n clearly orange (hence clearly not red), and successive patches pairwise visually

¹³ *VbD*, pp. 307–8.

¹⁴ The right-left direction requires: (1) if A entails C , then $v(C|A) = 1$; (2) if $A_1, \dots, A_n$ entail C , then $\neg C$ entails $\neg A_1 \vee \dots \vee \neg A_n$. The left-right direction requires: (3) if it's possible that $A_1, \dots, A_n$ are true and C false, then it's possible that $v(A_1) = 1, \dots, v(A_n) = 1$ and $v(C) = 0$. Now, (3) is pretty close to what Edgington's trying to show: it's an instance of the contrapositive of the left-right direction of $\text{Valid} \equiv \text{VCP}$. Representing verities as proportions of sharpenings in a supervaluationist setting may, however, allow us to justify (3); it's verified by the one sharpening model in which $A_1, \dots, A_n$ are all true and C false; that model exists whenever $A_1, \dots, A_n$ can be true with C false.

¹⁵ If the modality is metaphysical modality, then modal validity plausibly extends classical validity. Given standard views about metaphysical modality, '2+4=6' and hence also ' $\exists x.y(x \neq y)$ ' are modal logical truths but not classical logical truths. Likewise, the argument from 'Dorothy exists' to 'Dorothy is a human being' is modally but not classically valid. The slightly modified modal notion of validity in Section 15.2.2 does not have this feature.

¹⁶ For more detail see *VbD*, pp. 307–12.

indiscriminable to unaided normal human observers under ideal conditions. The sorites paradox arises because each of the following appears true:

‘ Ra_1 ’ is clearly true.

‘ Ra_n ’ is clearly false.

Each material conditional of the following form (where $1 \leq i \leq n$) is extremely plausible, and its negation extremely implausible:

$$(Si) \quad Ra_i \rightarrow Ra_{i+1}$$

The following argument is valid:

$$(A) \quad Ra_1, (S1), (S2), \dots, (Sn); \text{ therefore } Ra_n$$

Edgington’s solution to the paradox emerges from the verity theoretic properties of ‘ Ra_1 ’, ..., ‘ Ra_n ’.

As i increases from 1 to n , we gradually leave the clearly red patches, progress through patches increasingly distant (both physically and semantically) from clearly red, and terminate with clearly not red patches. Somewhere around the middle of the series, increasing i gradually decreases $v(Ra_i)$ from 1 to 0. The corresponding (Si) ’s have consequents with marginally lower verity than their antecedents; call them *dropping* (Si) ’s. On Edgington’s semantics, non-dropping (Si) ’s have verity 1, and dropping (Si) ’s have verity only marginally less than 1. We are therefore right to find each (Si) highly plausible: each is either clearly true or very close to it.

Given that argument (A) is valid, **Valid** $\equiv$ **VCP** implies: the unverity of ‘ Ra_n ’ cannot exceed the sum of the unverities of ‘ Ra_1 ’ and the (Si) ’s. ‘ Ra_1 ’ and the non-dropping (Si) ’s all have verity 1. What about the dropping (Si) ’s? Suppose verity decreases uniformly from 1 to 0 over m steps. Then there are m dropping (Si) ’s, each with verity $1 - \frac{1}{m}$. So their total unverity is $m \times \frac{1}{m} = 1$.¹⁷ **Valid** $\equiv$ **VCP** thus permits ‘ Ra_n ’ to have unverity 1. That is, ‘ Ra_n ’ can have verity 0 despite following by valid argument from premisses with high individual verities. The validity of argument (A) is therefore compatible with (a) ‘ Ra_1 ’ being clearly true, (b) ‘ Ra_n ’ being clearly false, and (c) each (Si) being clearly true or very close thereto. The features that generate the paradox are thereby rendered consistent. Moreover, when the premisses of a valid argument have low total unverity, **Valid** $\equiv$ **VCP** requires that the conclusion have high verity. So although valid reasoning does permit decrease in verity, belief in the conclusion on the basis of the premisses is not threatened. Provided the premisses have low total unverity, the conclusion will be both close to clearly true and far from clearly false. Adding that conclusion to one’s stock of beliefs will not significantly increase the distance from clear truth of one’s total doxastic state. Classical reasoning in the presence of vagueness is thereby vindicated.

¹⁷ Dropping the uniformity assumption blocks calculation of the verities of individual dropping (Si) ’s, though their total unverity remains $m \times \frac{1}{m} = 1$. It is in calculating the unverities of the dropping (Si) ’s that Edgington relies on the procedure for calculating conditional verities discussed in Section 15.1.1.

15.2 Verities as Truth-Values

What are verities? In light of their probabilistic structure, one might identify them with credences: ‘the credence that a person with no relevant ignorance other than about the precise line, would give to a statement like “that’s red”’.¹⁸ Against this proposal, Edgington argues that verity and credence occupy different theoretical roles; they interact with preferences in different ways.¹⁹

The natural alternative sees verities as a semantic phenomenon: verities are new truth-values. The classical semantic principle of bivalence says that every sentence has exactly one of the two classical truth-values. So if verities are truth-values, bivalence fails. Now, the classically valid Law of Excluded Middle (LEM) licenses assertion of $\lceil A \vee \neg A \rceil$ under any suppositions whatsoever, for any sentence A . Instances of bivalence follow from instances of LEM and the corresponding material conditionals $\lceil A \rightarrow a \text{ is true} \rceil$ and $\lceil \neg A \rightarrow a \text{ is false} \rceil$ (where a denotes A). So if verities are truth-values, those conditionals cannot all be true, given Edgington’s retention of classical logic. However, Edgington offers an interesting way of avoiding this conclusion, by denying that verities and classical truth-values are competitors. She attempts to reconcile her probabilistic semantics with classical semantics by suggesting that those theories are different descriptions of the same underlying phenomenon. This section examines and rejects that suggestion.

To simplify presentation, I use $\lceil BA \rceil$ to abbreviate $\lceil \neg DA \wedge \neg D\neg A \rceil$ (i.e. the formal counterpart of ‘It’s borderline whether A ’). I also extend the object-language with a truth-predicate ‘ T ’ (restricted somehow to avoid contradiction) and quotation-names for its own sentences. In the interests of readability, I omit object-language quotation marks so that, e.g., $\lceil T(p \wedge q) \rceil$ becomes $\lceil Tp \wedge q \rceil$. As usual, falsity is identified with having a true negation. I abbreviate $\lceil T(\neg A) \rceil$ as $\lceil FA \rceil$. I ignore meaningless and non-truth-evaluable sentences. I also abstract away from contextual effects, writing as if all sentences have contents and truth-values *simpliciter*, rather than only declarative sentences as used in particular contexts to make assertions.

15.2.1 Edgington on degrees of truth

The thesis under examination is:

$V = TV$ Verities are truth-values.

Edgington appears to reject $V = TV$, by arguing that verities are not degrees of truth.²⁰ This section shows that her argument is no threat to $V = TV$.

Here’s how Edgington argues that verities are not degrees of truth:

The degrees of truth argument Suppose it’s clearly borderline whether p : DBp . Then it isn’t clearly wrong to call ‘ p ’ true: $\neg D\neg Tp$. But it is clearly wrong to say that ‘ p ’ has verity 1, that it’s clearly true: $D\neg Dp$. So verity 1 is not truth. So verities are not degrees of truth.

¹⁸ *VbD*, p. 312.

¹⁹ *VbD*, pp. 312–15.

²⁰ *VbD*, p. 299.

This argument relies on two assumptions that are worth noting before continuing. Firstly, that clear borderline status is a coherent notion. Secondly, that when ' p ' is clearly borderline, ' Tp ' is also clearly borderline. The justification is presumably that when ' p ' is borderline, so is ' Tp '. Both assumptions might be questioned, though I won't do so here; the second is examined in more detail later (Section 15.3.2).

The final step in the argument relies on the assumption that if verities are degrees of truth, then maximum verity is truth. On that picture, intermediate verities are degrees of closeness to truth, the greater the degree, the closer to truth. Although that is a version of $V = TV$, it is not the only one.

The degrees of truth argument shows at most that truth is not maximum verity. A parallel argument shows that falsity is not minimum verity. The argument thus threatens one way of locating classical semantic concepts within the framework of verities, by identifying truth and falsity with maximum and minimum verity. Since $V = TV$ does not entail any identities between verities and classical truth-values, however, the degrees of truth argument does not threaten $V = TV$ itself.

$V = TV$ is in tension with Edgington's claim that 'I do not see verity as disturbing or displacing the concept of truth.'²¹ By truth, Edgington appears to mean classical bivalent truth. $V = TV$ replaces the two classical truth-values with many. Surely that counts as displacing the classical picture. However, another of Edgington's remarks suggests a resolution of this tension:

[W]e have two ways of classifying the same phenomena, each vague, and only vaguely related to each other. The [classical] way is simple, adequate when vagueness is not at issue, and too basic to be tampered with. The [verity] way is needed when vagueness is in focus: it enables us, we shall see, to solve the sorites paradox, in the context of a general theory of reasoning from vague premisses.²²

Edgington is endorsing:

Classification Classical semantics and verity semantics are two ways of classifying the same phenomenon.

The following two sections evaluate and reject this interesting proposal. Section 15.2.2 offers an account of the phenomenon in question that is neutral between classical and verity semantics. This account will give substance to the notion of a truth-value, thus fleshing out $V = TV$. Section 15.2.3 then argues against *Classification*. It follows that the classical theory and verity theory cannot be non-competing descriptions of semantics and truth-values in the sense of Section 15.2.2. The goal is to show that at most one of those theories captures the word-world relations that underwrite validity.

15.2.2 Two ways of classifying what?

What is classical semantics a theory of? This section attempts to answer without prejudice for or against classical semantics. The short answer is: semantics is the theory of truth-value assignments. More detail follows. This detail will be important

²¹ Ibid.

²² Ibid.

later, since I'll be distinguishing two conceptions of truth in Section 15.3.2, arguing that only one is relevant to semantics. Given a conception of semantics as the study of truth-conditions, the issue would be terminological: two notions of truth would yield two notions of semantics. The substantive conception of semantics, truth and the relations between them provided in this section hopefully alleviates this threat of merely verbal dispute.

Semantics is the study of truth-value assignments. Truth-value assignments are ways of distributing truth-values across sentences permitted by the meanings of the connectives. Which connectives? That depends on our interests. Sometimes, our narrow interests may cover only narrowly logical connectives. On other occasions, our more inclusive interests may encompass, say, the '□' of necessity, or the 'D' of clarity, as when we study modality and vagueness. Our interests may even extend beyond connectives, to predicates like 'T' too. Whatever vocabulary concerns us, the goal is to capture the totality of truth-value assignments that its meanings permit, and to use that totality to characterise validity. Semantic theories will differ over how the meanings of various expressions constrain truth-value assignments, and over exactly how truth-value assignments relate to validity. For present purposes, we can set those differences aside. The difference between classical semantics and verity semantics that now concerns us lies in the truth-values themselves.

Classical semantics admits exactly two truth-values. Given $V = TV$, verity semantics admits many truth-values. What exactly are the truth-values about which these views disagree? They are relational properties of sentences, whose possession is jointly determined by: (i) the sentence's content; (ii) the state of reality.²³ I'll ultimately be arguing (in Section 15.3.2) that the best version of Edgington's view will take verities but not classical truth and falsity to be truth-values in this sense. I'll say a few words about content now before returning to truth-values.

The content of a sentence is the way it represents things as being: a sentence's content is that p iff it represents things as being such that p .²⁴ The analysis of representation is a deep and difficult matter that cannot be addressed here. It is, however, an adequacy condition on theories of content and vagueness that they ultimately mesh. Many-valued semanticists owe an account of representation that comports with their multiplicity of truth-values.

Classical semantics admits two content-determined ways for sentences to stand to reality. First way (truth): things are the way the sentence represents them as being. Second way (falsity): things are not the way the sentence represents them as being. The logical form of these locutions is a delicate matter. Classical semanticists prepared to avail themselves of quantification into sentence position have a ready answer:²⁵

Truth: $\lambda x[\exists P(x \text{ represents things as being such that } P, \text{ and } P)]$.

Falsity: $\lambda x[\exists P(x \text{ represents things as being such that } P, \text{ and not-}P)]$.

²³ Cook, 2009.

²⁴ Better: iff speakers of the appropriate linguistic community use the sentence to represent things as being such that p .

²⁵ Künne, 2003 uses quantification into sentence position in a similar analysis of linguistic truth.

The variable '*P*' occupies sentence-position here; it is not a metalinguistic nominal variable ranging over sentences. Many-valued semanticists owe a similarly detailed account of their multiplicity of truth-values and of the *closer to clearly true than* ordering on them.

The most natural way to explain the many-valued semanticist's multiplicity of truth-values is in terms of the two classical ones, as degrees to which sentences possess the classical truth-values. One sentence's being closer to clearly true than another can then be explained as the former's being true to a greater degree than the latter. Two points are worth noting. Firstly, on this view, intermediate verities are intermediate degrees of truth, and maximum verity is truth. This conflicts with Edgington's degrees of truth argument. As Sections 15.3.1 and 15.3.2 will show, however, that argument is not decisive. Secondly, this strategy requires the notion of degrees to which things are the way they are represented as being. It is not immediately obvious whether that is intelligible. An adequacy condition on attempts to explicate it is that things can be the way they are represented as being to only an intermediate degree, without thereby being inaccurately represented; for otherwise intermediate truth-values will merely be varieties of falsity. The most developed theory of degrees of truth that I know of, due to Nicholas J. J. Smith, meets this adequacy condition.²⁶

On Smith's view, an atomic sentence '*Fa*' represents an object, denoted by '*a*', as having a property, denoted by '*F*'. Vagueness arises from a mismatch between the structures of representation and represented: instantiation comes in degrees, which are absent from the representation. Degrees of instantiation are used to analyse truth-values: '*Fa*' has truth-value *n* iff the denotation of '*a*' has the denotation of '*F*' to degree *n*. However, this simply relocates the explanatory burden to the theory of instantiation. Because degrees of instantiation are no more readily understood than degrees of truth, using the former to explicate the latter provides little explanatory advance. The availability of a fully satisfying account of degrees of truth remains an open question.

An important additional question concerns the nature of the differences between truth-value assignments.²⁷ On one view, different assignments result from differences in meaning. Another view invokes differences in what the world described by the language is like. Hybrids are also available. In line with Edgington's modal conception of validity, however, the second view will be assumed here. On this approach, different truth-value assignments result from different possible states of reality.

I said that semantics is the study of how the meanings of the logical constants—or any other privileged vocabulary—constrain truth-value assignments. Different truth-value assignments result from different possible states of reality. So semantics is the study of how the logical operations expressed by the logical constants—more generally: the semantic values of privileged vocabulary—constrain the possible states of reality, and how those constraints are reflected in the possible assignments of truth-

²⁶ Smith, 2008, pp. 201–13.

²⁷ See Etchemendy, 1990 for discussion.

values to sentences. The relevant notion of possibility will be more permissive than metaphysical possibility is typically taken to be. So far as the meanings of the logical constants are concerned, it's possible that $2 + 2 \neq 4$. The impossibility here lies with the numbers two, four, and the *addition* operation, not with *conjunction*, *negation*, or any other logical operation. Taking different truth-value assignments to result from different metaphysical possibilities would therefore impose constraints on truth-value assignments beyond those resulting from the meanings of the logical connectives alone, contrary to the present notion of semantics. Isolating the contribution of the meanings of the connectives requires a space of logical possibilities more inclusive than the metaphysical possibilities. Different choices of logical connective will deliver different spaces of logical possibilities.

This section has left many questions open. My goal, however, was not to settle every issue. My goal was to articulate the subject-matter of semantics in a manner amenable to both classical and many-valued semanticists, so that we can understand and evaluate *Classification* and $V = TV$. On the view proposed, semantics is the study of word-world relations, and the constraints imposed on such relations by the meanings of the object language logical constants (or whatever non-logical vocabulary concerns us). This conception of semantics and its connection with truth-value assignments places a substantive constraint on the nature of truth: truth-values are relational properties that sentences possess by virtue of their content and the state of reality. Section 15.3.2 will discuss a conception of truth that doesn't meet this constraint. Other notions of semantics may well be available, though they won't concern me here.

15.2.3 Against Classification

This section argues that *Classification* is false. Central classical semantic principles fail, if classical truth-values and verities provide two ways of classifying the same system of word-world relations. This leaves it open whether verities or classical truth and falsity are truth-values in the sense just outlined. Section 15.3 attempts to settle the matter in favour of verities.

How do assignments of verities and classical truth-values interact? Two plausible constraints on the relationship between a language's classical state c and verity state v are:

Upwards closure of truth Necessarily, for any sentences A, B , if $c(A) = \text{True}$ and $v(B) \geq v(A)$, then $c(B) = \text{True}$.

Downwards closure of falsity Necessarily, for any sentences A, B , if $c(A) = \text{False}$ and $v(B) \leq v(A)$, then $c(B) = \text{False}$.

Anything closer to clear truth (falsity) than a truth (falsheood) is also true (false). The classical truth-values are exclusive. So we can define a mapping π from verities to truth-values such that, for all i :

$\pi(i) = \text{True}$ iff, for some A : $v(A) = i$ and $c(A) = \text{True}$.

$\pi(i) = \text{False}$ iff, for some A : $v(A) = i$ and $c(A) = \text{False}$.

Consider the *T-set* $\{i : \pi(i) = \text{True}\}$ and the *F-set* $\{i : \pi(i) = \text{False}\}$. The closure principles ensure that every member of the F-set is less than every member of the

T-set. So those sets are disjoint. Given the following two assumptions, every verity belongs to the T-set or the F-set. (1) The classical semantic principle of bivalence: every sentence possesses some classical truth-value. (2) Every verity is possessed by some sentence. Since there are fewer sentences than verities, assumption (2) is problematic. The problem can be avoided by considering assignments of verity to open sentences, relative to assignments of values to variables.²⁸ Assuming that every possible shade between clear red and clear orange is actually instantiated, the open sentence ' Rx ' possesses each verity relative to some assignment of value to ' x '.

We've just seen that the T-set and the F-set partition the verities, with the members of the former all greater than those of the latter. So the greatest lower bound of the T-set belongs to one or other of those sets. Let's assume it belongs to the F-set.²⁹ Say that a *truth-threshold* is any verity t that satisfies:

$$c(A) = \text{True} \text{ iff } v(A) > t; \text{ and } c(A) = \text{False} \text{ iff } v(A) \leq t.$$

Truth is verity greater than the truth-threshold t . The greatest lower bound of the T-set is a truth-threshold. So the closure principles imply:

Threshold There is a truth threshold.

The argument for *Threshold* relies only on the exclusivity of classical truth and falsity, the closure principles, and assumptions (1) and (2) above.³⁰ Only (1) is controversial in the present context. But (1) is bivalence, hence part of classical semantics. So *Classification* implies *Threshold*. I'll now argue that *Threshold* is incompatible with classical semantics, and hence that *Classification* is false.

As far as the argument for *Threshold* goes, each verity assignment may be compatible with many classical assignments. Bivalence and the closure principles imply that a truth threshold exists without settling its location. None of the arguments below assume that a language's verity state uniquely settles its classical state.

One may interpret Edgington as rejecting *Threshold* when she says:

A sufficient condition for the truth of A is that $v(A) = 1$; a sufficient condition for the falsity of A is that $v(A) = 0$. Beyond that, there is no precise mapping.³¹

However, there is a complication with this interpretation of Edgington. She only explicitly rejects the existence of a precise mapping from verities to classical truth-values. Taking 'precise' to mean non-vague, this is consistent with the existence of a vague such mapping: for some verities, it's vague whether π maps them to True or to False. This fits Edgington's earlier quoted suggestion that verity assignments and

²⁸ The right hand side of the first biconditional governing π becomes: iff, for some A and for some variable assignment α , $v(A, \alpha) = i$ and $c(A, \alpha) = \text{True}$. The second biconditional requires a similar modification.

²⁹ Nothing turns on this: if the assumption is false, redefine 'truth-threshold' by substituting ' $\geq$ ' and ' $<$ ' for ' $>$ ' and ' $\leq$ ' respectively.

³⁰ The argument also requires the verity-theoretic analogue of bivalence: every sentence possesses some verity. Otherwise, the truths (or falsehoods, or both) may outstrip the sentences with any given kind of verity. I take this assumption to be uncontroversial in the present context.

³¹ *VbD*, p. 299.

classical assignments are both ‘vague, and only vaguely related to each other’.³² Properly evaluating this suggestion would require a theory of vague mappings. Fortunately, we needn’t provide one here. The arguments below assume only that a truth threshold exists, not that its location is precise. Any (total) mapping from verities to truth-values that respects the closure principles verifies *Threshold*. So the following arguments show that Edgington should deny the existence of any mapping whatsoever, whether vague or precise.

THE CLOSURE ARGUMENT

Threshold implies that truth isn’t closed under conjunction and falsity isn’t closed under disjunction. In classical semantics, truth is closed under conjunction and falsity is closed under disjunction. So *Threshold* is incompatible with classical semantics.

To see why *Threshold* implies that truth isn’t closed under conjunction, suppose ‘*p*’ and ‘*q*’ are independent. Then $v(q | p) = v(q)$. So:

$$v(p \wedge q) = v(p) \times v(q | p) = v(p) \times v(q).$$

Suppose:

The truth threshold $t = 0.7$.

$$v(p) = v(q) = 0.75.$$

Then:

$$v(p \wedge q) = 0.75 \times 0.75 = 0.56.$$

Since $v(p)$ and $v(q)$ are both above the threshold, ‘*p*’ and ‘*q*’ are both true. Since $v(p \wedge q)$ is below the threshold, ‘*p* $\wedge$ *q*’ is not true but false. So truth isn’t closed under conjunction, and a false conjunction has true conjuncts. This counterexample to the modal validity of conjunction introduction also undermines Edgington’s claim to preserve classical logic.

This problem doesn’t arise if $t = 1$. However, that generates three other problems. Firstly, it conflicts with Edgington’s degrees of truth argument in Section 15.2.1. Secondly, all verities other than 1 become varieties of falsity. Thirdly, falsity cannot be identified with having a true negation: when $0 < v(p) < 1 = t$, neither ‘*p*’ nor ‘ $\neg p$ ’ exceeds the truth-threshold; so if bivalence holds, ‘*p*’ is false despite ‘ $\neg p$ ’ being untrue.

To see why falsity isn’t closed under disjunction, suppose ‘*p*’ and ‘*q*’ are independent. Then:

$$v(p \vee q) = v(p) + v(q) - v(p \wedge q) = v(p) + v(q) - (v(p) \times v(q)).$$

Suppose:

The truth threshold $t = 0.7$.

³² Ibid.

$$v(p) = v(q) = 0.65.$$

Then:

$$v(p \vee q) = 0.65 + 0.65 - 0.65^2 = 1.3 - 0.42 = 0.88.$$

Since $v(p)$ and $v(q)$ are both below the threshold, ' p ' and ' q ' are both false. Since $v(p \vee q)$ is above the threshold ' $p \vee q$ ' is not false but true. So falsity isn't closed under disjunction, and a true disjunction has false disjuncts.

THE T-IMPLICATION ARGUMENT

Threshold invalidates the argument from ' p ' to ' Tp '. Classical semantics validates that argument. So if *Threshold* is true, classical semantics is false.

To show that *Threshold* invalidates the argument from ' p ' to ' Tp ', we need a semantic clause for ' T '. The natural candidate is:

$$v(\mathsf{Tp}) = 1 \text{ iff } v(p) > t; v(\mathsf{Tp}) = 0 \text{ otherwise.}$$

This ensures that ' T ' serves as an object-language device for reporting the presence of what the meta-theory regards as truth: ' Tp ' is clearly true when ' p ' exceeds the truth threshold, and clearly false otherwise. Now, suppose $0 < v(p) < t$. By the clause just given: $v(\mathsf{Tp}) = 0$. So ' Tp ' has maximum unverity: $1 - v(\mathsf{Tp}) = 1$. But ' p ' does not have maximum unverity: $1 - v(p) < 1$. So the unverity of ' Tp ' is greater than the unverity of ' p '. So the argument from ' p ' to ' Tp ' lacks the VCP, and, by *Valid* $\equiv$ VCP, is therefore invalid on Edgington's semantics.

One might object to this clause for ' T ' on the grounds that the boundary between truth and untruth is vague, and hence that $v(\mathsf{Tp})$ shouldn't sharply shift from 0 to 1 when $v(p)$ exceeds t . The natural implementation of this idea allows $v(\mathsf{Tp})$ to increase from 0 to 1 gradually, as $v(p)$ approaches and exceeds t . So suppose $v(\mathsf{Tp})$ begins to increase from 0 when $v(p)$ exceeds $t - \epsilon$. Then the argument goes through as before, if we vary the initial supposition to: $0 < v(p) < t - \epsilon$. Nothing has been gained.

Why does classical semantics validate the argument from ' p ' to ' Tp '? Modal validity, recall, is necessary truth-preservation. The point of ' T ' is to report the presence of that which valid arguments preserve. So ' Tp ' will have that feature whenever ' p ' does; that is, the argument from ' p ' to ' Tp ' is truth-preserving. Since that rests on no contingent assumptions, the argument from ' p ' to ' Tp ' is necessarily truth-preserving.

Note how radical a view this is, even by the standard of non-classical semantics. Supervaluation semantics has familiarised us with the idea that a material conditional ' $p \rightarrow \mathsf{Tp}$ ' might not be true.³³ However, that's not because supervaluation allows ' p ' to be true without ' Tp ' being true. Rather, supervaluation makes the whole conditional neither true nor false when ' p ' is neither true nor false. Supervaluation thus validates the argument from ' p ' to ' Tp '. Supervaluationists may therefore retain the biconditional ' p iff Tp ' by interpreting 'iff' as mutual entailment. The T-implication argument shows that defenders of *Threshold* cannot. I'll discuss this kind of issue in more detail later (Section 15.3.2).

³³ The *locus classicus* is Fine, 1975.

These arguments expose the radical consequences of *Threshold*. The closure argument shows that central classical semantic principles fail. The T-implication argument shows that the logical properties of truth are radically non-classical. So *Threshold* is incompatible with classical semantics, hence also with *Classification*. Defenders of *Classification* must reject the closure principles from which *Threshold* follows. That, I submit, is a deeply unattractive view. Although it allows one to regard bivalent truth and verity as ways of classifying the same phenomenon, it is obscure why one would wish to do so, given the non-classical conceptions of truth and semantics that result.

The option remains of regarding classical semantics as a limiting case of verity semantics, in the sense of being perfectly adequate when borderline cases are ignored. The second and third sentences in the quote from *VbD*, p. 299 (p. 247 of this volume) might be interpreted as suggesting this kind of view. Yet this is an unsatisfying defence of classical semantics and *Classification*. It amounts to an admission that borderline cases are counterexamples to bivalence. This is not a way for the classical theory and the verity theory to be non-competing descriptions of a single semantic phenomenon.

Edgington is guided by a parallel between credence and the lottery paradox on the one hand, and vagueness and the sorites on the other. An analogue of *Classification* for credence and belief is available: credence and belief co-classify the same doxastic phenomenon.³⁴ Analogues of *Threshold* and the closure argument fit quite nicely with that view. Arguably, lotteries show that rational belief isn't closed under conjunction: one can rationally believe, of each ticket-holder, that he won't win, without believing the conjunction of those beliefs. The lesson is that Edgington's guiding parallel can only take us so far.³⁵

15.3 Vagueness as *Sui Generis*?

I've argued that *Classification* is false: the classical theory and verity theory are not non-competing descriptions of the same semantic phenomenon. At most one of those theories captures the semantics of vague language. In more recent work, Edgington also rejects *Classification*. On her new view, however, the theory of verities is not a semantic theory at all.³⁶ This section examines and rejects this proposal.

Drawing on a proposal of David Barnett's, Edgington suggests that vagueness is *sui generis*, something that cannot 'be illuminatingly understood as a species of some more general phenomenon'.³⁷ Call this view primitivism. Primitivism is incompatible with $V = TV$, the thesis that verities are truth-values in the sense of Section 15.2.1; for $V = TV$ makes borderline status a species of the more general phenomenon of semantic evaluation.

³⁴ Edgington, 1997, p. 299 endorses that thesis.

³⁵ Rosanna Keefe's contribution to this volume discusses the putative parallel between prefaces, lotteries and vagueness in more detail.

³⁶ Edgington, MS, §§1–4; Edgington, 2010, §5.6.

³⁷ Edgington, MS, §1, footnote 1, MS, p. 2; Barnett, 2009.

Edgington retains classical logic. So borderline status is compatible with LEM: if it's borderline whether p , then p or not- p , it's just not clear which. Edgington retains bivalence in a similar manner: if it's borderline whether p , then ' p ' is true or false, it's just not clear which. This is just what primitivism should lead us to expect; for if vagueness isn't a semantic phenomenon, then borderline status shouldn't interfere with classical semantic evaluation, or with the classical logic it validates. Primitivism thus provides a way to make classical logic and semantics compatible with borderlines status.

Despite preserving classical semantics and logic, primitivism is methodologically unattractive. We should be content to take X as primitive only when two conditions are satisfied. Firstly, we can use X to provide satisfying accounts of other phenomena. That, for example, is why knowledge is plausibly taken as primitive.³⁸ Nothing similar applies here; clarity's explanatory connections beyond the theory of vagueness are thin at best. Secondly, no satisfying analysis of X should be available. Since $V = TV$ promises one such analysis, we should be content with primitivism only if $V = TV$ is problematic.

Edgington and Barnett present two arguments that appear to threaten $V = TV$. Both arguments purport to show that bivalence should hold without restriction, even in borderline cases. It appears to follow that we should reject $V = TV$ because it makes the multitude of verities incompatible with bivalence. This section shows that these arguments against $V = TV$ fail. The problem is that bivalent truth and falsity needn't be truth-values in the sense of Section 15.2.2; they might play some other, non-semantic, theoretical role. Indeed, I will argue that Edgington and Barnett's arguments for bivalence presuppose that they do.

Section 15.3.1 outlines Edgington and Barnett's arguments for bivalence. The first argument is invalid unless supplemented by a substantive principle about truth. That principle also provides the best justification for one of the second argument's key inferences. Section 15.3.2 argues that this principle is incompatible with the conceptions of semantics and truth outlined in Section 15.2.2. Rather than presenting a problem for $V = TV$, Edgington and Barnett's arguments for bivalence require a non-semantic conception of bivalent truth. An appropriate such conception is outlined that satisfies the principle underwriting the arguments for bivalence.

Primitivism and the view I'll be defending are structurally similar: bivalent truth and vagueness are fundamentally different phenomena. This renders the existence of borderline cases compatible with bivalence. The views differ over the theoretical roles of verities and classical truth. The primitivist says: vagueness and verities are *sui generis*, to be understood in their own terms or not at all; semantic evaluation involves classical truth.³⁹ The view I'll be defending says: verities are truth-values, i.e. relational properties possessed by virtue of content and the state of reality; classical truth plays a merely expressive, non-semantic role.

³⁸ Williamson, 2000; Hossack, 2007.

³⁹ Primitivism could in principle be combined with the non-semantic, expressive conception of classical truth outlined in Section 15.3.2. An account is then required of the truth-values in the sense of Section 15.2.1. However, the end of Section 15.3.2 claims that classical semanticists are in the privileged position of allowing the expressive and semantic roles to coincide.

15.3.1 *Bivalence at the border*

Although Edgington does not discuss $V = TV$ directly, she does argue against views on which borderline cases present counterexamples to bivalence. $V = TV$ replaces the two classical truth-values with many, assigning intermediate values to borderline cases. So on $V = TV$, borderline cases are counterexamples to bivalence. Edgington's arguments thus appear to threaten $V = TV$. These arguments also appear to motivate primitivism, by showing that borderline status shouldn't be understood as a semantic phenomenon that interferes with bivalence. I'll argue that these appearances are illusory. Edgington's arguments do show something interesting, but not something that conflicts with $V = TV$.⁴⁰

First argument for borderline bivalence ' $p \vee \neg p$ ' is classically valid, hence valid on Edgington's view. So if the following material conditionals are true, so is the corresponding instance of bivalence:

$$(T \rightarrow) \quad p \rightarrow Tp \quad \neg p \rightarrow Fp$$

So if ' p ' is neither true nor false, one of those conditionals is untrue. Edgington rejects this because 'it follows that either [p] but it's not true that [p], or [$\neg p$] but it's not false that [p].'⁴¹ But 'what more could it take to make it true that [p] than [p]?'⁴²

Second argument for borderline bivalence Suppose it's clearly borderline whether p : DBp . If borderline cases are counterexamples to bivalence, then ' p ' clearly isn't true: $D\neg Tp$. But 'this is in tension with our ambivalence about the question whether [p]' is true, our temptation to see it as "sort of" true'.⁴³ When it's clearly borderline whether p , it should be clearly borderline whether ' p ' is true, not clear that it isn't. So when it's borderline whether p , ' p ' should be borderline true, rather than untrue. Likewise for $\neg p$ and falsity. So borderline cases shouldn't be counterexamples to bivalence.

In both arguments, the choice of ' p ' was arbitrary. So we can generalise to bivalence. What exactly do these arguments show?

The first argument is invalid in some many-valued settings. In classical semantics, untrue material conditionals are false; their antecedents are true and their consequents false. Yet in a many-valued setting, an untrue material conditional may not be false. Many-valued semanticists can allow untrue ' $p \rightarrow Tp$ ' without true ' $p \wedge \neg Tp$ ', by denying that untruth of a material conditional requires true antecedent plus false consequent; untrue antecedent with false consequent may suffice. This is what happens in standard supervaluationist semantics, where (super)truth-valueless antecedent plus (super>false consequent suffices for (super)truth-valuelessness of the whole conditional. Although ' $p \vee \neg p$ ' is (globally) valid, the corresponding instance of bivalence can fail because ' $p \rightarrow Tp$ ' is neither (super>true nor (super>false when ' p '

⁴⁰ The arguments considered in the text are from Edgington, 2010, p. 104 and Edgington, MS, §2; versions are also found in Barnett, 2009, pp. 14–16.

⁴¹ Edgington, 2010, p. 104.

⁴² Barnett, 2009, p. 15.

⁴³ Edgington, 2010, p. 104.

is neither (super)true nor (super>false.⁴⁴ As Edgington and others have noted, verities can be modelled by proportions of sharpenings.⁴⁵ So this problem with Edgington's first argument for borderline bivalence afflicts her verity-theoretic setting too.

Can we fix the argument? Not simply by replacing the material conditionals ($T \rightarrow$) with logical implications. Supervaluation semantics validates the arguments from ' p ' to ' Tp ' and from ' $\neg p$ ' to ' Fp ', though bivalence still fails. The reason is that (global) validity is necessary preservation of (super)truth, and ' $p \vee \neg p$ ' can be (super)true when neither disjunct is; both ' Tp ' and ' Fp ' are (super>false in such cases, and so (T) lack (super)-truth-value. This is consistent with ' Tp ' being (super)true whenever ' p ' is, and likewise for ' Fp ' and ' $\neg p$ '. On this view, no more is required for ' p ' to be true than that p .

Of course, one might simply insist on the material conditionals ($T \rightarrow$). That amounts to treating them as logical truths in the following sense: the meanings of the logical connectives and ' T ' (and quotation names) preclude truth-value assignments on which ($T \rightarrow$) are untrue. Since ' $p \vee \neg p$ ' is logically true, the corresponding instance of bivalence is too. Generalising: every instance of bivalence is a logical truth. Indeed, instances of bivalence are inter-derivable with (corresponding sentences of the form) ($T \rightarrow$) given LEM and the uncontroversial converses of (those sentences of the form) ($T \rightarrow$).

Treating every sentence of the form ($T \rightarrow$) as valid embodies a substantive commitment about truth. The converse conditionals are uncontroversially valid. So every instance of these material biconditionals is valid too:

$$(T \leftrightarrow) \quad A \leftrightarrow TA \quad \neg A \leftrightarrow FA$$

The material biconditional is a test for identity of truth-value. If ' $A \leftrightarrow B$ ' is logically true, it's logically impossible for A and B to receive different truth-values; the meanings of the logical constants alone preclude states of reality to which the contents of A and B stand in different ways. Now, if ' T ' isn't counted a logical constant, instances of ($T \leftrightarrow$) are not logically true: so far as *conjunction*, *negation*, etc. are concerned, the atomic predication ' Tp ' may have a different truth-value from ' p '. On the present proposal, the meaning of ' T ' (and quotation names) suffices to exclude such states. The meaning of ' T ' is therefore captured by some function f from sentences to sentence-contents such that, for any sentence A and possibility w : $f(A)$ and the content of A stand to w in the same way. That's the minimum required for the meaning of ' T ' to preclude truth-value assignments that differentiate between A and ' TA '. What function could f be? There appears to be only one candidate: $f(A) =$ the content of A . Unless that identity holds, it's mysterious what prevents states of reality from inducing truth-value assignments that falsify instances of ($T \leftrightarrow$). So on the present proposal, that treats the instances of ($T \leftrightarrow$) as logical truths, the meaning of ' T ' guarantees:

Identity For any sentence A , the content of ' TA ' is the same as that of A .

⁴⁴ I'm assuming that ' Tp ' is true on a sharpening iff ' p ' is supertrue—i.e. true on all sharpenings—and hence that ' Tp ' is superfalse when ' p ' is anything other than supertrue.

⁴⁵ Edgington, 1997, pp. 315–16; McGee and McLaughlin, 1994, §5. Note that Edgington sees the parallel with supervaluation semantics as formally but not philosophically/metaphysically helpful.

What we have just seen is that a fixed up version of the first argument for borderline bivalence, one that's not invalid in the semantic setting Edgington's attacking, presupposes *Identity*.⁴⁶

Consider the second argument for borderline bivalence. It assumes that clear borderline cases are possible. This is not trivial, but I won't question it here. The argument also assumes that when it's borderline whether p , we are (or perhaps should be) ambivalent about whether ' p ' is true. Why should we grant that? I see two candidate reasons.

The first takes it as part of the data to be accommodated, part of our ordinary linguistic behaviour that semantic theory should capture. That may be questioned. When we take it to be borderline whether p , we certainly are ambivalent, confused and conflicted about whether p . This confusion about whether p might lead ordinary speakers to express similar attitudes about whether ' p ' is true. We should not assume without argument, however, that ordinary speakers are right to do so, or that this is relevant to truth. Our interest is in the interaction of borderline status with truth, not with ordinary speakers' usage of 'true'. Ordinary uses of 'true' may not always express truth. Ordinary speakers are also often mistaken, about even the most commonplace of concepts. And truth is not the most commonplace of concepts. Truth has a complex theoretical role to play, one aspect of which is its place in semantic theory. The fundamental concept of semantic theory should not be held hostage to ordinary usage of 'true'.

A better reason to think that it being borderline whether p entails borderline status in ' Tp ' is this: ' p ' and ' Tp ' are always intersubstitutable *salve veritate*. Given this assumption, ' Bp ' implies ' BTp ', and hence that the confusion and ambivalence characteristic of borderline status is appropriately directed towards ' Tp ' whenever appropriately directed towards ' p '.

Unrestricted substitutability entails *Identity*. One way to see this is by substituting ' Tp ' into one side of the logically true material biconditional ' $p \leftrightarrow p$ ', to obtain a logically true instance ' $p \leftrightarrow \text{Tp}$ ' of ' $(\text{T} \leftrightarrow)$ '. Since ' p ' was arbitrary, we can generalize to the logical truth of every instance of ' $(\text{T} \leftrightarrow)$ '. As argued above, that implies *Identity*. Alternatively, let $\ominus$ be a content-identity connective: ' $p \ominus q$ ' is true iff ' p ' has the same content as ' q '. Since ' $p \ominus p$ ' is true, intersubstitutability implies that ' $p \ominus \text{Tp}$ ' is true. Since ' p ' was arbitrary, we can generalise to *Identity*.

In sum, the first argument for borderline bivalence is unpersuasive because invalid in the many-valued settings it is supposed to refute. Fixing the argument leads to *Identity*. *Identity* also underwrites the second argument for borderline bivalence. Consider too the degrees of truth argument in Section 15.2.1. The key inference there is from ' Bp ' ($= \neg \text{Dp} \wedge \neg \text{D}\neg p$) to ' $\neg \text{D}\neg \text{Tp}$ '. Appeal to *Identity* is the most natural way to justify that. Rejecting *Identity* therefore allows identification of truth and falsity with verities 1 and 0 respectively, and of intermediate verities with intermediate degrees of truth.

⁴⁶ For an interesting response to a similar argument for *Identity*, see Asher et al., 2009, §3.2; Dever, 2009, esp. §§9, 12.

This diagnosis of *Identity*'s place in Edgington's thought is reinforced by:

There is no reason to deny the equivalence of 'It is true that A' and 'A', or of 'It is false that A' and '¬A'. If $v(A) = 0.5$, $v(\text{It is true that } A) = 0.5$. As $v(A \text{ or not } A) = 1$, so $v(\text{It is true that } A \text{ or it is false that } A) = 1$, even if each disjunct gets 0.5. . . . A principle which is stronger than bivalence is rejected: the principle that every proposition is either [clearly] true or [clearly] false, that every proposition has verity 1 or 0. We saw earlier: a disjunction can be [clearly] true, without either disjunct being [clearly] true.⁴⁷

Commitment to *Identity* manifests here in the equivalence of 'It is true that A' with A. What should many-valued semanticists, and defenders of $V = TV$ in particular, make of Edgington's use of *Identity*? The next section argues that they should take it to show that, on Edgington's view, classical truth and falsity don't concern the word-world relations that underwrite validity; their theoretical role differs from that of truth-values in the sense of Section 15.2.2.

15.3.2 Two roles for truth

Edgington's arguments for the compatibility of borderline status with bivalence presuppose *Identity*. It might therefore appear that many-valued semanticists should reject *Identity*. This section argues that the appearance is misleading. A different moral can be drawn instead: two theoretical roles associated with truth come apart.⁴⁸ I'll begin by describing these roles.

A truth-predicate allows linguistic expression of contents it would otherwise be impossible or impractical to express. Suppose I know that Keith has uncannily good judgement, so good in fact that he's incapable of error. Without a truth-predicate (or equivalent), the closest I can come to expressing this is to begin uttering an infinite conjunction: if Keith says that whales are fish, then whales are fish, and if Keith says that grass is blue, then grass is blue, and if Keith says that necessity is *a priority*, then necessity is *a priority*, . . . Since the conjunction is infinite, however, I cannot finish expressing it. So I cannot communicate just how good Keith's judgement is. But it may be important for me to do so. If Dorothy is going to the races with Keith tomorrow, where she will be gambling with my money, it matters greatly to me whether she will bet the same way as Keith. A truth-predicate allows me to overcome this difficulty by wrapping up the infinitely many conjuncts into a finite universal generalization:

For any sentence x that Keith uses tomorrow to make an assertion, x as then used by Keith will be true.

A truth-predicate thus serves an expressive function; it allows linguistic communication of otherwise inexpressible contents. Call this first theoretical role the *expressive role* for truth.⁴⁹ A predicate F cannot occupy the expressive role unless: for any

⁴⁷ *VbD*, p. 310.

⁴⁸ McGee and McLaughlin, 1994, 2000 and McGee, 2005 advocate a similar view.

⁴⁹ For discussion of the expressive utility of truth, see Azzouni, 2004, ch. 1. Speakers of languages that permit quantification into sentence position have no need for a truth-predicate. Even if such quantification is intelligible, however, it does not appear to be present in natural language: we need an expressive truth-predicate.

content-bearer A , ' Fy ' is guaranteed to have the same content as A under an assignment of A to ' y '. So the expressive role for truth requires *Identity*.

Truth's second theoretical role lies in empirical semantics. It is a non-trivial matter what kind of worldly circumstances suffice for the truth of 'the sky is blue'. That expression could have been used in many different ways, to express many different contents. The way a concrete (or syntactically individuated) string represents things as being is contingent, determined largely by its use within a linguistic community. One aspect of empirical semantics is the association of expressions with worldly items and circumstances—i.e. with contents—on the basis of contingent features of use. One central goal of this enterprise is to delineate the worldly circumstances under which sentences are true (as they are used within the relevant linguistic community) on the basis of the meanings of their sub-sentential components plus syntactic structure. Truth is thus involved in a minimal standard of correctness for empirical semantics.⁵⁰ Call this the *semantic role* for truth.

It is not a priori that a single notion occupies both the expressive and semantic roles. The expressive role requires a certain type of predicate and arises from a practical need. The semantic role concerns the underlying structure of the language to which the expressive truth-predicate belongs, and the word-world relations involving expressions of the language, as determined by use. It should not be assumed without argument that an expressive truth-predicate fills the needs of empirical semantics. Because the arguments for borderline bivalence presuppose *Identity*, they target expressive truth most directly. We can therefore retain $V = TV$ without faulting those arguments, by taking $V = TV$ to concern semantic truth. Indeed, we should take it to do so anyway. The point of an expressive truth-predicate is not to classify sentences on the basis of how their content stands to reality (or a possible state thereof), but to form a sentence with the same content as another sentence A from an expression that denotes A . To the extent that an expressive truth-predicate does effect such a classification, it does so parasitically on the word-world relations arising from the contents of the sentences classified. Explicating those relations is the purpose of semantic truth. And that way of classifying sentences—on the basis of how their content stands to reality—is what I used in Section 15.2.2 to characterize the truth-value assignments that underwrite validity. Differentiating these roles for truth allows bivalent expressive truth to cohabit with many-valued semantic truth.

Why connect validity with semantic truth? Because an argument is valid when a certain relation holds between its premisses and conclusion at every possibility. To make sense of that, we need relations between sentences and such possibilities. These sentence-possibility relations are, in effect, what Section 15.2.2 used to explicate truth-values. They result from the association of strings with contents that underwrites semantic truth. One might reply that an expressive truth-predicate allows us to introduce appropriate sentence-possibility relations, by evaluating ' Tp ' relative to such possibilities (holding fixed the actual meaning of ' p '). However, insofar as an expressive truth-predicate allows us to simulate sentence-possibility relations, it does

⁵⁰ Another truth-involving standard of correctness is that the semantics get entailment relations correct.

so parasitically on the contents of the sentences themselves, and the relations between sentences and possibilities induced by those contents. When 'T' is an expressive truth-predicate, the evaluability of ' $\text{T}p$ ' relative to a possibility w depends on the evaluability of the concrete string ' p ' relative to w , which itself depends on (a) ' p ' representing things as being a certain way, and (b) w adjudicating whether things are that way. The evaluability of ' $\text{T}p$ ' relative to w thus relies on a semantic notion of truth.

Another way in which this parasitism emerges is via the fact that an expressive truth-predicate does not require its satisfiers to be any distinctive way. An expressive truth predication ' $\text{T}p$ ' is (typically) not even about ' p '; it is about whatever ' p ' itself is about. The evaluability of ' $\text{T}p$ ' relative to w therefore rests on the evaluability of ' p ' relative to w . Since ' p ' is an intrinsically meaningless concrete string, its evaluability relative to w relies on its being used in a contentful manner, which is just what semantic truth captures. Since validity involves evaluating sentences relative to possibilities, validity requires semantic truth.

An independent argument shows that predicates for the many-valued semanticist's truth-values will not serve the expressive role. Let ' F ' be an object-language predicate for reporting the presence of verity n . Then ' Fp ' should have maximum verity when $\nu(p) = n$ and minimum verity otherwise. But with an expressive truth-predicate ' T ': $\nu(\text{T}p) = \nu(p)$. We should therefore expect divergence between an expressive truth-predicate and the linguistic resources in which many-valued semantics is couched. Only in a two-valued setting should we expect these to coincide, where bivalent semantic truth-conditions exhaust content. The separation of expressive and semantic truth is no threat to many-valued semantics.

On this approach, classical bivalent truth plays no role in empirical semantics, or in the analysis of validity. It exists only to fill an expressive need. If LEM holds without restriction, then so does one form of bivalence: expressive bivalence. Instances of expressive bivalence are merely syntactic variants on instances of LEM: their content is the same. Expressive bivalence thus places no constraints on the semantics of vagueness, beyond validation of LEM. In constructing such a semantics, one may use as many truth-values as one likes.

Edgington's arguments for borderline bivalence do not refute $V = TV$. They do, however, force advocates of $V = TV$ to distinguish the semantic and expressive roles for truth. The question remains whether verities or classical truth-values should occupy the semantic role. Given the unattractiveness of primitivism about vagueness, I conclude by suggesting that the best version of Edgington's view will treat verities as semantic truth-values, hence concerned with word-world relations, whereas classical bivalent truth serves a merely expressive role.

References

- Asher, N., Dever, J., and Pappas, C. (2009) Supervaluations debugged. *Mind* 118(472): 901–33.
 Azzouni, J. (2004) *Deflating Existential Consequence: A Case for Nominalism*. Oxford: Oxford University Press.
 Barnett, D. (2009) Is vagueness sui generis? *Australasian Journal of Philosophy*, 87: 5–34.

- Cook, R. T. (2009) What is a truth value and how many are there? *Studia Logica* 92(2): 183–201.
- Dever, J. (2009) The disunity of truth. In R. J. Stainton and C. Viger (eds), *Compositionality, Context and Semantic Values: Essays in Honour of Ernie Lepore*, chapter 7. Netherlands: Springer.
- Edgington, D. (1997) Vagueness by degrees. In R. Keefe and P. Smith (eds), *Vagueness: A Reader*, Cambridge, MA: MIT Press, chapter 16.
- Edgington, D. (2010) Sorensen on vagueness and contradiction. In R. Dietz and Moruzzi, S. (eds.), *Cuts and Clouds: Vagueness, Its Nature and Its Logic*, chapter 5. Oxford: Oxford University Press.
- Edgington, D. (MS) Vagueness: filling the gaps. Unpublished.
- Etchemendy, J. (1990) *The Concept of Logical Consequence*. Cambridge, MA and London: Harvard University Press.
- Fine, K. (1975) Vagueness, truth and logic. *Synthese* 30: 265–300.
- Hossack, K. (2007) *The Metaphysics of Knowledge*. Oxford: Oxford University Press.
- Keefe, R. (2000) *Theories of Vagueness*. Cambridge: Cambridge University Press.
- Künne, W. (2003) *Conceptions of Truth*. Oxford: Oxford University Press.
- McGee, V. (2005) Two conceptions of truth?—comment. *Philosophical Studies* 124(1): 71–104.
- McGee, V. and McLaughlin, B. (1994) Distinctions without a difference. *The Southern Journal of Philosophy* 33: 203–61.
- McGee, V. and McLaughlin, B. (2000) The lessons of the many. *Philosophical Topics* 28: 129–51.
- Shapiro, S. (2006) *Vagueness in Context*. Oxford: Clarendon Press.
- Smith, N. J. J. (2008) *Vagueness and Degrees of Truth*. Oxford: Oxford University Press.
- Williamson, T. (1994) *Vagueness*. London, New York: Routledge.
- Williamson, T. (2000) *Knowledge and Its Limits*. Oxford: Oxford University Press.

Bibliography of works by Dorothy Edgington

- Edgington, Dorothy (1976) 'The Applicability of Bayesian Convergence-of-Opinion Theorems to the Case of Actual Scientific Inference' (with J. Dorling), *British Journal for the Philosophy of Science*, 27: 2160–1.
- Edgington, Dorothy (1980–81) 'Meaning, Realism and Bivalence', *Proceedings of the Aristotelian Society*, LXXX: 153–73.
- Edgington, Dorothy (1985) 'The Paradox of Knowability', *Mind*, 94: 557–68.
- Edgington, Dorothy (1985) 'Verificationism and the Manifestation of Meaning', *Proceedings of the Aristotelian Society Supplementary Volume* LIX: 33–52.
- Edgington, Dorothy (1986) 'Do Conditionals Have Truth Conditions?', *Crítica*, XVIII: 3–39, reprinted in R. I. G. Hughes (ed.), *A Philosophical Companion to First Order Logic*. Indianapolis: Hackett, 1993, pp. 28–49.
- Edgington, Dorothy (1988) 'Un argumento de Orayen en favor del condicional material', *Revista Latinoamericana de Filosofía*, 13: 54–8.
- Edgington, Dorothy (1990) 'Explanation, Causation and Laws', *Crítica*, XXII: 55–73.
- Edgington, Dorothy (1991) 'Do Conditionals Have Truth Conditions?' (expanded version) in F. Jackson (ed.), *Conditionals*. Oxford: Oxford Readings in Philosophy, pp. 176–201.
- Edgington, Dorothy (1991) 'The Mystery of the Missing Matter of Fact', *Proceedings of the Aristotelian Society Supplementary Volume* LXV: 185–209.
- Edgington, Dorothy (1992) 'Changing Beliefs Rationally: Some Puzzles' in J. Ezquerro and J. M. Larrazabel (eds), *Cognition, Semantics and Philosophy: Proceedings of the First International Colloquium on Cognitive Science*. Dordrecht: Kluwer, pp. 47–73.
- Edgington, Dorothy (1992) 'Validity, Uncertainty and Vagueness', *Analysis* 52: 193–204, reprinted in D. Graff and T. Williamson (eds), *Vagueness*. Ashworth, 2002.
- Edgington, Dorothy (1993) 'Wright and Sainsbury on Higher-order Vagueness', *Analysis* 53, 193–200, reprinted in D. Graff and T. Williamson (eds), *Vagueness*. Ashworth, 2002.
- Edgington, Dorothy (1994) Contributions to T. Honderich (ed.), *The Oxford Companion to Philosophy* Oxford: Oxford University Press: 'Assertion', p. 62; 'Intuitionism, mathematical', p. 415.
- Edgington, Dorothy (1995) 'Conditionals and the Ramsey Test', *Proceedings of the Aristotelian Society Supplementary Volume*, LXIX: 67–86.
- Edgington, Dorothy (1995) 'On Conditionals', *Mind*, 104: 235–329.
- Edgington, Dorothy (1995) 'The Logic of Uncertainty', *Crítica*, 27: 27–54.
- Edgington, Dorothy (1996) 'Lowe on Conditional Probability', *Mind*, 105: 617–30.
- Edgington, Dorothy (1996) 'Vagueness by Degrees' in Rosanna Keefe and Peter Smith (eds), *Vagueness: A Reader*. Cambridge MA: MIT Press, pp. 294–316.
- Edgington, Dorothy (1997) 'A Commentary' in M. Woods, *Conditionals*. Oxford: Clarendon Press, pp. 95–137.
- Edgington, Dorothy (1997) 'Mellor on Chance and Causation', *British Journal for the Philosophy of Science*, 48: 411–33.
- Edgington, Dorothy (1997) 'Truth, Objectivity, Counterfactuals and Gibbard', *Mind*, 106: 108–16.
- Edgington, Dorothy (1998) 'Modal Logic', in D. Edgington, S. Guttenplan, and M. Machover, *Symbolic Logic: A Guide*, University of London, pp. 77–95.

- Edgington, Dorothy (1998) 'Williamson on Iterated Attitudes', *Proceedings of the British Academy*, Book 95: *Philosophical Logic*, ed. T. Smiley, pp. 135–58.
- Edgington, Dorothy (2000) 'General Conditionals: A Response to Kölbel', *Mind*, 109: 109–16.
- Edgington, Dorothy (2001) 'Conditionals', in L. Goble (ed.), *Blackwell Guide to Philosophical Logic*, Oxford: Blackwell, pp. 385–414.
- Edgington, Dorothy (2001) 'Conditionals' in the *Stanford Encyclopedia of Philosophy*, Stanford: CSLI. <http://www.plato.stanford.edu/entries/conditionals/> (23 pages).
- Edgington, Dorothy (2001) 'Indeterminacy *de re*', *Philosophical Topics* 28(1): 27–44.
- Edgington, Dorothy (2001) 'The Philosophical Problem of Vagueness', *Legal Theory*, 7: 371–8.
- Edgington, Dorothy (2002) 'Williamson on Vagueness, Identity and Leibniz's Law' in A. Bottani, M. Carrara and P. Giaretta (eds.), *Individuals, Essence and Identity: Themes in Analytic Metaphysics*. Dordrecht: Kluwer.
- Edgington, Dorothy (2003) 'What if? Questions about Conditionals'. *Mind and Language*, October, 380–401.
- Edgington, Dorothy (2004) 'Counterfactuals and the Benefit of Hindsight', in P. Dowe and P. Noordhof (eds), *Chance and Cause*. Routledge: International Library of Philosophy, pp. 12–27.
- Edgington, Dorothy (2004) 'Two Kinds of Possibility'. *Proceedings of the Aristotelian Society Supplementary Volume*, pp. 1–22.
- Edgington, Dorothy (2005) 'Ramsey's Legacies on Conditionals and Truth', in D. H. Mellor (ed.), *Ramsey's Legacy*. Oxford: Oxford University Press, pp. 37–52.
- Edgington, Dorothy (2005) 'The Mystery of the Missing Boundary', *Philosophy and Phenomenological Research*, LXXI (3): 704–11.
- Edgington, Dorothy (2006) 'Pragmatics of the Logical Constants', in, E. LePore and B. C. Smith (eds), *The Oxford Handbook on Philosophy of Language*. Oxford: Oxford University Press, pp. 768–93.
- Edgington, Dorothy (2007) 'On Conditionals', in D. Gabbay and E. Guenther (eds), *Handbook of Philosophical Logic*, second edition, Volume 14, Dordrecht: Springer, pp. 127–222. (This a slightly expanded version of 'On Conditionals', *Mind*, 1995. The main additions being in the final section on counterfactuals.)
- Edgington, Dorothy (2008) 'Counterfactuals', *Proceedings of the Aristotelian Society*, 1–21.
- Edgington, Dorothy (2009) 'Conditionals, Truth and Assertion', in I. Ravenscroft (ed.), *Minds, Ethics and Conditionals: Themes from the Philosophy of Frank Jackson*. Oxford: Oxford University Press, pp. 283–310.
- Edgington, Dorothy (2010) 'Possible Knowledge of Unknown Truth', *Synthese*, 173(1): 41–52.
- Edgington, Dorothy (2010) 'Sorensen on Vagueness and Contradiction', in R. Dietz and S. Moruzzi (eds), *Cuts and Clouds: Vagueness, its Nature and its Logic*. Oxford: Oxford University Press, pp. 91–106.
- Edgington, Dorothy (2011) 'Causation First: Why Causation is Prior to Counterfactuals', in C. Hoerl, T. McCormack, and S. Beck (eds), *Understanding Counterfactuals/Understanding Causation*. Oxford: Oxford University Press, pp. 230–41.
- Edgington, Dorothy (2011) 'Conditionals, Causation and Decision'. *Analytic Philosophy* 52(2): 75–87.
- Edgington, Dorothy (2014) 'Estimating Conditional Chances and Evaluating Counterfactuals', *Studia Logica* 102(4): 691–707.
- Edgington, Dorothy (2016) 'Schiffer on Vagueness, Indeterminacy and Conditionals', in Gary Ostertag (ed.), *Meanings and Other Things: Themes from the Work of Stephen Schiffer*. Oxford: Oxford University Press.
- Edgington, Dorothy (2018) 'Counterfactuals', in O. Bueno and S. Shalkowski (eds), *Routledge Handbook of Modality* (forthcoming).

Reviews include

Adams, E. (1978) *The Logic of Conditionals*, *Mind*, 87: 619–23.

Forbes, G. (1988) *The Metaphysics of Modality*, *Philosophical Quarterly*, 38: 152 365–70.

Mackie, J. L. (1976) *Truth, Probability and Paradox*, *Mind*, 85: 303–8.

In preparation

Edgington, Dorothy (2014) *Towards Reasonable Doubt*, on probability, possibility, conditionals and vagueness, for Oxford University Press.

Index of Names

- Adams, Ernest 13ff., 20, 30–1, 58, 62, 64, 67, 72,
104–6, 109–11, 117, 214, 216
- Bennett, Jonathan 11–13, 16, 19, 25, 100
- de Finetti, Bruno 10, 13, 25n, Chapter 5
passim
- Dummett, Michael 10–12, 15, 69, 196n, 203
- Edgington, Dorothy Chapter 1 passim,
Chapter 2 passim, Chapter 3 passim, 40–2,
50, Chapter 5 passim, 76, 79, Chapter 7
passim, 122, 128n, 138, 141n, 164, 174, 190,
Chapter 12 passim, Chapter 13 passim, 227,
237, Chapter 15 passim
- Égré Paul 19n, 30n, 41, 52, 53n, 228, 232, 235
- Fine, Kit 151n, 152n, 154, 201n, 222n, 241n,
253n
- von Fintel, Kai 19n, 22n, 30n, 33–4, 36n,
Chapter 4 passim, 94n, 126, 129, 130n, 135n,
138, 167–8
- Frege, Gottlob 9–11, 15. *See also* Frege-Geach
problem
- Gibbard, Allan 3, 32, 101, 106
- Gillies, Anthony 23n, 50, 200n
- Goodman, Nelson 11, 42, 76
- Hájek, Alan 5–6, 18, 30n, 105n
- Heim, Irene 19, 22ff
- Higginbotham, James 32–3, 40ff, 209n
- Iatridou, Sabine 33–4, Chapter 4 passim, 100n,
Chapter 8 passim, 168
- Kamp, Hans 11, 224n, 235
- Karttunen, Lauri 21, 78–80, 86, 93n
- Keefe, Rosanna 6, 215n, 219n, 222n, 223n,
224n, 242n, 254n
- Kratzer, Angelika 3, 4, 18, Chapter 3 passim,
40, 91, 137, 165, 167, 168, 172
- Lewis, David 2, Chapter 2 passim, Chapter 3
passim, 40, 66, 67, 76, 85, 86, 91, 100ff., 114,
190–1, 198n, 200–2, 224n, 237
- McGee, Vann 3, 45, 110–12, 257n, 259n
- Moss, Sarah 18, 35n, 50, 208n, 234
- Pollock, John Chapter 11 passim
- Raffman, Diana 228, 232, 235
- Ramsey, Frank 9, 10, 13, 58, 61, 65ff
- Rothschild, Daniel 4, 19n, 25n, 30n, 33n, 35n
- Smith, Nicholas 219n, 220, 221, 249
- Stalnaker, Robert 2, 11, 13, 14, 18, 20, 22,
26, 43–5, 46, 54, 55, 66, 68, 79, 82, 85,
86, 94, 96, 102n, 103n, 104n, 105n, 108n,
200–1
- Swanson, Eric 27n, 29, 33
- Van Fraassen, Bas 2, 15, 103n, 105, 106,
145n, 149
- Williamson, Timothy 2n, 6–7, 15, 43n, 45–7,
171n, 196–7, 199n, 200n, 201n, 203, 204,
208n, 220n, 229n, 241n, 255n
- Yalcin, Seth 24n, 29, 31, 33n, 36

Index of Topics

- Adams's Thesis 2, 3, 30, 31, 60, 64, 66, 67,
Chapter 7 passim, 104–6, 109–11, 117
- Belief Revision 15, 37, 70–2, 105, 190–2
- Bivalence 5, 12, 23n, 28, 42, 46, 239ff, 246–7,
251ff
- Choice Functions 115ff
- Church-Fitch Paradox 6, 15, Chapter 12 passim
- Common Ground 44, 82, 84, 104
- Conditional Assertion 2, 28, 29, 61, 63,
Chapter 7 passim
- Conditional Excluded Middle 34, 43, 45–7
- Conditional Degree of Belief Chapter 7 passim
- Conditional Probability 1, 13, 17, 30, 35,
Chapter 5 passim, 76, Chapter 7 passim,
171n, 207, 208n
- Conditional Credence, *see* Conditional Degree
of Belief
- Conditionals:
 - Biscuit Conditionals 130n
 - Counterfactuals 2, 7, 11, 13, 14, 17, 19n, 26n,
46, 50, 64, 65, 66, Chapter 6 passim,
Chapter 7 passim, Chapter 8 passim, 142,
151n, 153–4, Chapter 12 passim
 - Embedded Conditionals 2–3, 22, 24–5, 32–6,
Chapter 4 passim, 209
 - Indicative Conditionals 2, 3n, 13, 14, 17, 19n,
20, 24, 26ff, 35, Chapter 4 passim,
Chapter 5 passim, 86n, 87n, 94n, 96–8,
Chapter 7 passim, 207
 - Material Conditional 19n, 20, 22, 26, 29–30,
34, 54, 59, 60, 62, 68, 196, 209, 216n, 242,
245, 246, 253, 256, 257
 - No Truth-Value 2–4, Chapter 3 passim,
100–2
 - Restrictor Analysis 4, Chapter 4 passim,
49–53, 137–8
 - Stalnaker Conditional 20, 22, 44, 45, 54, 55,
Chapter 7 passim
 - Strict Conditionals 20, 22, 26, 34, 36, 200
 - Subjunctive Conditionals, *see*
Counterfactuals
 - Variably Strict Conditionals 19n, 20, 26, 34
- Contraposition 13, 45–7
- Credence, *see* Degree of Belief
- Degree of Belief 4–5, 66–7, 70–1, 100, 190, 212,
214–19, 221–2, 225, 230
- Epistemic Modals 27, 29, 36, 101–6,
Chapter 10 passim
- The Equation, *see* Adams's Thesis
- Expressivism 2, 32, 101–6
- Fitch's Paradox, *see* Church-Fitch
Paradox
- Fake Past Tense 127–9
- Frege-Geach Problem 2, 7, 32, 209
- Implicatures 49, 78, 80n, 82n
- Impossible Worlds Chapter 9 passim
- Indeterminacy 101, 114–15, 200
- Knowability Paradox, *see* Church-Fitch
Paradox
- Law of Excluded Middle 12, 103n, 246, 255,
257, 261
- Lottery Paradox 13, 14, 212, 227, 231, 235, 237,
239, 254
- Modal Completion 154, 160–1
- Modus Ponens 3, 4, 45–7, 59, 213,
220, 221
- Polarity Chapter 6 passim
- Preface Paradox 4–6, 13, Chapter 13 passim,
Chapter 14 passim, 254n
- Presupposition 29n, 42, 61, 64, Chapter 6
passim, 102, 106, 113, 116, 130, 131, 132,
180, 182, 184, 185, 188
- Scalar assertions 81–3, 85, 90, 92, 94,
97–8
- Trivalence 23, 25, 27–9
- Triviality Result 2, 30n, 66, 105, 106
- Vagueness 4–6, 14, 17, 18, 29n, 165n, 171n,
Chapter 13 passim, Chapter 14 passim,
Chapter 15 passim
- Degree theoretic 5, 17, 212, 214,
215–19, 220ff, Chapter 15 passim
- Epistemicism 229n, 231n
- Supervaluationism 6, 218, 222–6, 244n, 253,
256, 257
- Verificationism 12. *See also* Church-Fitch
Paradox