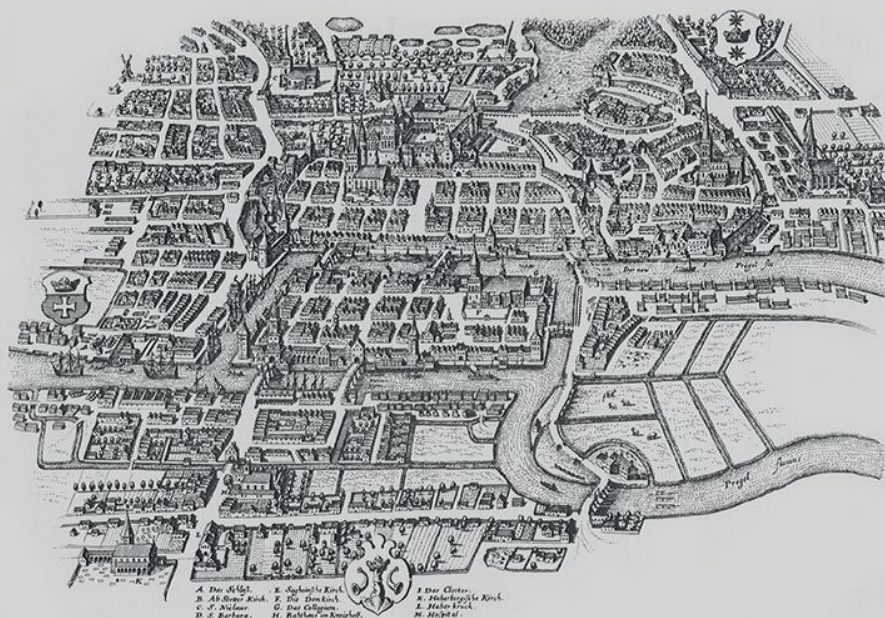


EDITED BY
ALEXANDER REUTLINGER AND JUHA SAATSI



EXPLANATION BEYOND CAUSATION

PHILOSOPHICAL PERSPECTIVES
ON NON-CAUSAL EXPLANATIONS

OXFORD

Explanation Beyond Causation

Explanation Beyond Causation

*Philosophical Perspectives on
Non-Causal Explanations*

EDITED BY

Alexander Reutlinger and Juha Saatsi

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 2018

The moral rights of the authors have been asserted

First Edition published in 2018

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: 2017963783

ISBN 978-0-19-877794-6

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

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 Figures</i>	vii
<i>Notes on Contributors</i>	ix

Introduction: Scientific Explanations Beyond Causation <i>Alexander Reutlinger and Juha Saatsi</i>	1
---	---

Part I. General Approaches

1. Because Without Cause: Scientific Explanations by Constraint <i>Marc Lange</i>	15
2. Accommodating Explanatory Pluralism <i>Christopher Pincock</i>	39
3. Eight Other Questions about Explanation <i>Angela Potochnik</i>	57
4. Extending the Counterfactual Theory of Explanation <i>Alexander Reutlinger</i>	74
5. The Mathematical Route to Causal Understanding <i>Michael Strevens</i>	96
6. Some Varieties of Non-Causal Explanation <i>James Woodward</i>	117

Part II. Case Studies from the Sciences

7. Searching for Non-Causal Explanations in a Sea of Causes <i>Alisa Bokulich</i>	141
8. The Development and Application of Efficient Coding Explanation in Neuroscience <i>Mazviita Chirimuuta</i>	164
9. Symmetries and Explanatory Dependencies in Physics <i>Steven French and Juha Saatsi</i>	185
10. The Non-Causal Character of Renormalization Group Explanations <i>Margaret Morrison</i>	206

Part III. Beyond the Sciences

11. Two Flavours of Mathematical Explanation	231
<i>Mark Colyvan, John Cusbert, and Kelvin McQueen</i>	
12. When Are Structural Equation Models Apt? Causation versus Grounding	250
<i>Lina Jansson</i>	
<i>Index</i>	267

List of Figures

1.1. Some grades of necessity	26
5.1. Königsberg's bridges	102
5.2. A Hamiltonian walk	110
7.1. A "sand sea": the Algodones dunes of SE California	149
7.2. A sequence of high-speed motion photographs of the processes of saltation and reptation	151
7.3. Examples of ripple defects	154
7.4. Subaqueous sand ripples on the ocean floor	158
8.1. Four kinds of explanation	167
8.2. Receptive fields of retinal ganglion cells	170
8.3. Visual illusions explained by lateral inhibition	171
8.4. Re-coding to reduce redundancy	174
9.1. A symmetrical triangle	187
9.2. Balance in equilibrium	187

Notes on Contributors

ALISA BOKULICH is Professor of Philosophy at Boston University and Director of the Center for Philosophy & History of Science, where she organizes the Boston Colloquium for Philosophy of Science. She is an Associate Member of Harvard University's History of Science Department and Series Editor for *Boston Studies in the Philosophy and History of Science*. She is the author of *Reexamining the Quantum-Classical Relation: Beyond Reductionism and Pluralism* (Cambridge University Press, 2008) and her research focuses on scientific models and explanations in the physical sciences, including the geosciences.

MAZVIITA CHIRIMUTA is Associate Professor of History and Philosophy of Science at the University of Pittsburgh. She received her PhD in visual neuroscience from the University of Cambridge in 2004, and held postdoctoral fellowships in philosophy at Monash University (2005–8) and Washington University St. Louis (2008–9). Her principal area of research is in the philosophy of neuroscience and perceptual psychology, and her book *Outside Color: Perceptual Science and the Puzzle of Color in Philosophy* was published by MIT Press in 2015.

MARK COLYVAN is Professor of Philosophy at the University of Sydney and a Visiting Professor at the Munich Center for Mathematical Philosophy at the Ludwig-Maximilians University in Munich. He holds a BSc (Hons) in mathematics (from the University of New England) and a PhD in philosophy (from the Australian National University). His main research interests are in the philosophy of mathematics, philosophy of logic, decision theory, risk analysis, and philosophy of ecology and conservation biology. He is the author of *The Indispensability of Mathematics* (Oxford University Press, 2001), *Ecological Orbits: How Planets Move and Populations Grow* (Oxford University Press, 2004, with co-author Lev Ginzburg), *An Introduction to the Philosophy of Mathematics* (Cambridge University Press, 2012), and numerous papers.

JOHN CUSBERT is a Research Fellow in Philosophy at the University of Oxford. He has a PhD in philosophy from the Australian National University. His research focuses on various topics in and around probability and decision theory, ethics, and metaphysics.

STEVEN FRENCH is Professor of Philosophy of Science at the University of Leeds. He is Co-Editor in Chief of the *British Journal for the Philosophy of Science* and Editor in Chief of the Palgrave Macmillan series New Directions in Philosophy of Science. His most recent book is *The Structure of the World: Metaphysics and Representation* (Oxford University Press, 2014) and his next one is *Applying Mathematics: Immersion, Inference, Interpretation* (with Otavio Bueno).

LINA JANSSON is Assistant Professor of Philosophy at the University of Nottingham. She received her PhD from the University of Michigan, Ann Arbor, and previously worked at Nanyang Technological University in Singapore. She works on issues related to explanation, laws of nature, and confirmation from within the history and philosophy of science, general philosophy of science, and the philosophy of physics. She has published work on issues related to non-causal explanation, explanations in Newton's *Principia*, ground, parsimony, and probability in Everettian quantum theories.

MARC LANGE is Theda Perdue Distinguished Professor and Philosophy Department Chair at the University of North Carolina at Chapel Hill. He is the author of *Because Without Cause: Non-Causal Explanation in Science and Mathematics* (Oxford University Press, 2016), *Laws and Lawmakers* (Oxford University Press, 2009), *An Introduction to the Philosophy of Physics: Locality, Fields, Energy, and Mass* (Blackwell, 2002), and *Natural Laws in Scientific Practice* (Oxford University Press, 2000).

KELVIN MCQUEEN is Assistant Professor of Philosophy and affiliate of the Institute for Quantum Studies at Chapman University. He has a PhD in philosophy from the Australian National University. He works on a variety of topics in the philosophy of science, the philosophy of physics, the philosophy of mind, and metaphysics.

MARGARET MORRISON is Professor of Philosophy at the University of Toronto and is a fellow of the Royal Society of Canada and the Leopoldina-German National Academy of Sciences. She received her PhD in philosophy of science from the University of Western Ontario. Her work covers a broad range of topics in the philosophy of science including physics and biology. Some of her publications include *Reconstructing Reality: Models, Mathematics and Simulation* (Oxford University Press, 2015), *Unifying Scientific Theories* (Cambridge University Press, 2000), and over sixty articles in various journals and edited collections.

CHRISTOPHER PINCOCK is Professor of Philosophy at the Ohio State University. He works on topics at the intersection of the philosophy of science and the philosophy of mathematics. He is the author of *Mathematics and Scientific Representation* (Oxford University Press, 2012).

ANGELA POTOCHNIK is Associate Professor at the University of Cincinnati. She earned her PhD at Stanford University. She works on a variety of topics in philosophy of science, including methodological issues in population biology, especially evolutionary and behavioral ecology; idealized models in biology and in science more generally; scientific explanation; relations among different projects and fields of science; how gender and social factors influence science; and the history of logical empiricism, especially the work of Otto Neurath. She is the author of, *Idealization and the Aims of Science* (University of Chicago Press, 2017).

ALEXANDER REUTLINGER is Assistant Professor at the Ludwig-Maximilians-Universität München (Munich Center for Mathematical Philosophy). He works on

topics in philosophy of science and neighbouring areas of epistemology and metaphysics (including topics such as explanation, causation, probabilities, *ceteris paribus* laws, idealizations, reduction, and models). He previously held positions as a postdoctoral research fellow at the University of Cologne and as a visiting fellow at the University of Pittsburgh's Center for Philosophy of Science.

JUHA SAATSI is Associate Professor at the University of Leeds. He works on various topics in philosophy of science, and he has particular interests in the philosophy of explanation and the scientific realism debate.

MICHAEL STREVEIS is Professor of Philosophy at New York University. He has written on scientific explanation, complexity, probability and probabilistic inference, causation, the social structure of science, and concepts of natural kinds and other theoretical concepts. He previously taught at Stanford University and Iowa State University.

JAMES WOODWARD is Distinguished Professor in the Department of History and Philosophy of Science at the University of Pittsburgh. Prior to 2010 he was the J. O. and Juliette Koepfli Professor at the California Institute of Technology. He is the author of *Making Things Happen: A Theory of Causal Explanation* (Oxford University Press, 2003) which won the 2005 Lakatos award, a past president of the Philosophy of Science Association (2010–12), and a fellow of the American Academy of Arts and Sciences.

Introduction

Scientific Explanations Beyond Causation

Alexander Reutlinger and Juha Saatsi

What is a scientific explanation? This has been a central question in philosophy of science at least since Hempel and Oppenheim's pivotal attempt at an answer in 1948 (also known as the covering-law model of explanation; Hempel 1965: chapter 10). It is no surprise that this question has retained its place at the heart of contemporary philosophy of science, given that it is one of the sciences' key aims to provide *explanations* of phenomena in the social and natural world around us. As philosophers of science, we naturally want to grasp and to explicate what exactly scientists are doing and aiming to achieve when they explain something.

In his classic *Four Decades of Scientific Explanation*, Salmon (1989) details the shift from Hempel and Oppenheim's "epoch-making" logical empiricist beginnings to a mixture of subsequent perspectives on scientific explanation involving ideas concerning causation, laws, theoretical unification, pragmatics, and statistics. Although Salmon believes that causal accounts of explanation (including his own version) are considerably successful, he ultimately advocated a pluralistic outlook. According to his pluralism, different approaches to explanation are worth pursuing and they should be understood as complementing one another rather than competing with each other. He articulates this pluralism, for instance, in his claim about the "peaceful coexistence" of causal and unificationist accounts.¹ According to Salmon, the four decades of intense philosophical activity on scientific explanation since 1948 did not result in anything like a consensus, and his prediction was that no broad consensus was likely to emerge after 1989, at least not in the short term.

However, Salmon's pluralist outlook and his portrayal of the history of the debate (articulated in his *Four Decades*) were largely lost in subsequent philosophical work. The two decades following the publication of Salmon's book in 1989 became the

¹ Salmon's well-known illustration of his pluralism is captured in the story of the friendly physicist (Salmon 1989: 183).

decades of causal accounts of explanation. As causal accounts came to dominate the philosophical scene, this tendency also resulted in establishing a research focus on causation itself, and since the late 1980s philosophers have made considerable progress in analysing various aspects of causation. For example, they have explicated different notions of causation, causal processes, causal mechanisms, and causal models, and they have achieved a better understanding of the connection between causes and different kinds of idealizations, of the link between causation and temporal order, and, indeed, of the kinds of explanations that causal information supports. According to causal accounts, the sciences explain by identifying the causes of the phenomenon to be explained—or, according to the mechanist version of causal accounts, by identifying the causal mechanisms for that phenomenon (for surveys see Andersen 2014; Woodward 2014).

Causal accounts have been considered to be attractive for several reasons. The focus on causal-mechanical aspects of explanation has undoubtedly been in many ways a good response to the shortcomings of the covering-law model (and of some alternative approaches to explanation). Moreover, the proponents of causal accounts have also taken a closer look at detailed case studies of real-life explanations in the sciences instead of merely analysing toy examples. The proponents of causal accounts have also advanced the field by taking seriously case studies from the life and social sciences, freeing the debate from a (formerly) widespread physics chauvinism. And, indeed, many paradigmatic explanations in the sciences rely on information about causes and mechanisms. Hence, philosophers focusing on causal explanation have achieved a great deal by studying this aspect of the explanatory practices of science. As a result, *today* hardly anyone denies the explanatory significance and epistemic value of causal-mechanistic information provided by the sciences.

The domination of the causal accounts has shaped the subsequent debate on scientific explanation in several respects: in how arguments have been perceived and evaluated; what the criteria for an adequate account of scientific explanation have been taken to be (for instance, everybody had to talk about flagpoles, for better or worse), and so on. This spirit of a ‘causal hegemony’ can easily be detected in extant survey papers (such as Woodward 2014; Craver and Tabery 2017),² also in influential works advocating a causal approach to scientific explanation (for instance, Woodward 2003; Craver 2007; Strevens 2008), and last but certainly not least in the tacit presumptions and ‘common knowledge’ one encounters at various conferences and workshops.

The state of the field after six long decades suggests that something close to a consensus was reached: scientific explanation is a matter of providing suitable information about causes of the explanandum phenomenon. However, over the past decade or so this consensus has come under increasing scrutiny and suspicion as philosophers have more widely begun to rethink the hegemony of causal-mechanist accounts.

² However, Woodward’s entry in the *Stanford Encyclopedia of Philosophy* remains open-minded about the possibility of non-causal explanations (Woodward 2014: §7.1).

There are important precedents to this recent development. Indeed, although causal accounts did indeed dominate the philosophical scene in the 1990s and the 2000s, they were far from being the only game in town. From early on, a number of authors have drawn attention to non-causal ways of explaining, in particular in relation to unificationist accounts (Friedman 1974; Kitcher 1984, 1989; Bartelborth 1996), pragmatic accounts (van Fraassen 1980, 1989; Achinstein 1983), analyses of asymptotic explanations in physics (Batterman 2000, 2002), statistical and geometrical explanations (Lipton 1991/2004; Nerlich 1979), and other specific examples from various scientific disciplines (for instance, Forge 1980, 1985; Sober 1983; Ruben 1990/2012; Frisch 1998; Hüttemann 2004).

Over the past few years, this resistance to the causal hegemony has burgeoned quickly, and the present volume demonstrates this turning of the tide. Looking at the current literature, one particularly striking recent development is the increasing interest in the *limits* of causal accounts of explanation. The guiding idea is that although causation is certainly part of the truth about scientific explanation, it is unlikely to be the full story. Following this idea, philosophers have begun to explore the hypothesis that explanations in science sometimes go *beyond causation*. For instance, there seem to be genuinely non-causal explanations whose explanatory resources go ‘beyond causation’ as these explanations do not work by way of truthfully representing the causes of the phenomenon to be explained. Other scientific explanations go ‘beyond causation’ in the sense that their explanatory assumptions do not tell us anything about the causal mechanisms involved. In this spirit, a number of philosophers have argued that the repertoire of explanatory strategies in the sciences is considerably richer than causal accounts suggest. (See Reutlinger 2017 for a detailed survey of the present debate on non-causal explanations.)

The motivation for this shift of focus to explanations that go ‘beyond causation’ is easy to appreciate: there are plenty of compelling, real-life examples of *non-causal* explanations that causal accounts of explanation seemingly fail to capture. To be more precise, the new development in the philosophy of *scientific* explanation is the increasing recognition of interesting and varied examples of non-causal explanations of *empirical phenomena* to be found across the natural and social sciences.

Unsurprisingly, physics is a fertile ground for such examples, ranging from explanations involving symmetries and inter-theoretic relations, to theoretically more abstract explanations that rely on, for instance, renormalization group techniques. Moreover, in the more fundamental domains of physical theorizing, it seems relatively easy to find explanations that seem non-causal—in the first blush at least. Perhaps this does not come as a surprise to those sympathetic to increasingly popular scepticism about causation as a fundamental metaphysical category in physics (originating in the work of Ernst Mach and Bertrand Russell among others; see, for instance, Mach 1905; Russell 1912/13; Scheibe 2007: chapter 7). Such causal ‘anti-foundationalism’ is a contested topic in its own right, of course, but perhaps the difficulty of interpreting fundamental physics in plain causal terms already indicates that explanations

in fundamental physics operate in terms that go beyond causation (Price 1996; Price and Corry 2007).

One need not plunge the depths of fundamental physics to find compelling instances of non-causal explanations, however. Various philosophers have suggested that there are other kinds of non-causal explanations in the life and social sciences, such as mathematical, statistical, computational, network, optimality, and equilibrium explanations. Moreover, some of the most popular examples in the philosophical literature—the present volume included—involve rather simple empirical set-ups of strawberries and bridge-crossings. Philosophers' love of toy examples is due to the fact that simple though such examples are, they are sufficiently instructive to challenge the philosophy of explanation centred around causal accounts, giving rise to fruitful engagement between competing philosophical analyses. For instance, what explains the fact that 23 strawberries cannot be distributed equally among 3 philosophers (cf. Chapter 1)? Is this explanation non-causal? Is it non-causal because it is mathematical? Is it mathematical in some *distinct* kind of way (in which familiar mathematized, and possibly causal, explanations in science are not)? As the essays in this volume demonstrate, thinking carefully about some exceedingly simple cases alongside real-life scientific explanations is not only fun, but philosophically profitable!³

Let us pause for a second. Surely, one might think, the existence of non-causal explanations is old news. After all, the empirical sciences are not the only epistemic project striving for explanations. Proofs in logic and pure mathematics are at least sometimes taken to be explanatory—and if so, then proofs explain in a non-causal way (see, for instance, Mancosu 2015). In metaphysical debates, too, one finds a straightforward appeal to non-causal explanations: for instance, if some fact *A* grounds another fact *B*, then *A* is taken to be non-causally explanatory of *B* (see, for instance, Bliss and Trogdon 2016). However, the fact that mathematicians, logicians, and metaphysicians sometimes explain in non-causal terms is an interesting and related topic but it is not *the crucial motivation* for questioning the hegemony of causal-mechanist accounts of explanations in the natural and social sciences.⁴ But even if non-causal explanations in logic, mathematics, and metaphysics do not motivate a challenge to causal hegemony in philosophy of science, it is certainly worth exploring the relationship between non-causal explanations in mathematics, logic, and metaphysics, on the one hand, and non-causal explanations in the natural and social sciences, on the other hand.

³ Action or teleological explanations are also often treated as a particular kind of non-causal explanation, as, for instance, von Wright (1971, 1974) argues. However, the allegedly non-causal character of action explanations is (infamously) controversial and has led to an extensive debate (see Davidson 1980 for a defence of a causal account of action explanations). We will bracket the debate on action explanations in this volume.

⁴ Although the existence of non-causal explanations internal to, for instance, pure mathematics and logic has long been recognized, detailed philosophical accounts of such explanations have been under-developed. The dominance of causal models of explanation in philosophy of science is partly to be blamed, since much of this work did not seem to be applicable or extendible to domains such as mathematics, where the notion of causation obviously does not apply.

Now, what would be an appropriate philosophical reaction to examples of non-causal explanations from the natural and social sciences? Let us canvass in the abstract three possible ‘big picture’ reactions:

1. causal reductionism,
2. explanatory pluralism, and
3. explanatory monism.

First, while some are happy to give up the hegemony of causal accounts of explanation and to welcome non-causal ways of explaining empirical phenomena, others feel less pressure to do so. Some philosophers—including some featured in this volume—take the *seeming* examples of non-causal explanations to rather point to the need for a more sophisticated account of causal explanation. If the seemingly non-causal explanations can ultimately be understood as causal explanations after all, perhaps non-causal explanations of empirical phenomena are indeed rare and exotic (if not wholly non-existent). The attraction of such *causal reductionism* about explanation, if indeed true, lies in the fundamental causal unity it finds underlying the *prima facie* disparate activity of scientific explanation. One and the same conceptual framework provides a pleasingly unified philosophical theory of explanation, if all explanations in science—including alleged examples of non-causal explanations—turn out to ultimately function by providing causal information. In other words, causal reductionists would like to maintain and to defend the hegemony of causal explanation (see, for instance, Lewis 1986; more recently Skow 2014, 2016).

Second, one way to deny such causal reductionism is to accept some kind of *explanatory pluralism*. Pluralists adopt, roughly put, the view that causal and non-causal explanations are different types of explanations that are covered by two (or more) distinct theories of explanation.⁵ The core idea of a pluralist response to the existence of examples of causal and non-causal explanations is that causal accounts of explanations have to be supplemented with further accounts of non-causal explanations (a view Salmon was attracted to, as pointed out above, see Salmon 1989; more recently Lange 2016).

Third, an alternative to explanatory pluralism is *explanatory monism*: the view that there is one single philosophical account capable of capturing both causal and non-causal explanations by virtue of some ‘common core’ that they share. To take an analogy, consider the way in which some theories of explanation (such as Hempel’s or Woodward’s) account for both deterministic and probabilistic (causal) explanations. In an analogous way, a monist holds that one theory of explanation may account for both causal and non-causal explanation. Unlike the causal reductionist, the monist does not deny the existence of non-causal explanations. Rather, a monist holds that causal and non-causal

⁵ This notion of explanatory pluralism has to be distinguished from another kind of pluralist (or relativist) attitude towards explanations, according to which one phenomenon has two (or more) explanations and these explanations are equally well suited for accounting for the phenomenon.

explanations share a feature that makes them explanatory (for a survey of different strategies to articulate monism, see Reutlinger 2017).

The ‘big picture’ issue emerging from these three reactions is whether causal reductionism, explanatory pluralism, or explanatory monism provides the best approach to thinking about the similarities and differences between various causal and (seemingly) non-causal explanations of empirical phenomena. However, this ‘big picture’ question is far from being the only one, and we predict that these debates are likely to continue in the foreseeable future due to a number of other outstanding questions such as the following ones:

- How can accounts of non-causal explanations overcome the problems troubling the covering-law model?
- What is the best way to distinguish between causal and non-causal explanations?
- Which different types of non-causal explanations can be found in the life and social sciences?
- Is it possible to extend accounts of non-causal explanation in the sciences to non-causal explanations in other ‘extra-scientific’ domains, such as metaphysics, pure mathematics, logic, and perhaps even to explanations in the moral domain?
- What should one make of the special connection that some non-causal explanations seem to bear to certain kinds of idealizations?
- What role does the pragmatics of explanation play in the non-causal case?
- What are the differences between non-causal and causal explanatory *reasoning*, from a psychological and epistemological perspective?
- What does scientific understanding amount to in the context of non-causal explanations?

Let us now turn to a preview of the volume, which divides into three parts.

Part I addresses issues regarding non-causal explanations from the perspective of *general philosophy of science*. By articulating suitable conceptual frameworks, and by drawing on examples from different scientific disciplines, the contributions to this part examine and discuss different notions of non-causal explanation and various philosophical accounts of explanation for capturing non-causal explanations.

Marc Lange presents a view that is part of a larger pluralist picture. For him, there is no general theory covering all non-causal explanations, let alone all causal and non-causal explanations taken together. But Lange argues that a broad class of non-causal explanations works by appealing to *constraints*, viz. modal facts involving a stronger degree of necessity than physical or causal laws. Lange offers an account of the order of explanatory priority in explanations by constraint, and uses it to distinguish different kinds of such explanations. He illustrates the account with paradigmatic examples drawn from the sciences.

Christopher Pincock probes different strategies for spelling out what pluralism—the view that, roughly put, explanations come in several distinct types—amounts to in relation to causal vs. non-causal explanations. He contrasts *ontic* vs. *epistemic* versions

of pluralism, and he finds room within both versions to make sense of explanatory pluralism in relation to three types of explanations: causal, abstract, and constitutive types of explanation. Moreover, he also draws attention to several problems that explanatory pluralism raises requiring further consideration and, thereby, setting a research agenda for philosophers working in a pluralist spirit.

Angela Potochnik argues that theories of explanation typically have a rather narrow focus on analysing explanatory dependence relations. However, Potochnik argues that there is no good reason for such a narrow focus, because there are many *other features* of explanatory practices that warrant philosophical attention, i.e., other features than the causal or non-causal nature of explanatory dependence relations. The purpose of Potochnik's contribution is mainly to convey to the reader that it is a serious mistake to ignore these 'other features'. She draws philosophical attention to features of explanations such as the connection between explanation and understanding, the psychology of explanation, the role of (levels of) representation for scientific explanation, and the connection between the aim of explanation and other aims of science. Her contribution is a plea for moving the debate beyond causal—and also beyond non-causal—dependence relations.

Alexander Reutlinger defends a monist approach to non-causal and causal explanations: the counterfactual theory of explanation. According to Reutlinger's counterfactual theory, both causal and non-causal explanations are explanatory by virtue of revealing counterfactual dependencies between the explanandum and the explanans (illustrated by five examples of non-causal scientific explanations). Moreover, he provides a 'Russellian' strategy for distinguishing between causal and non-causal explanations within the framework of the counterfactual theory of explanation. Reutlinger bases this distinction on 'Russellian' criteria that are often associated with causal relations (including causal asymmetry, time asymmetry, and distinctness).

Michael Strevens proposes to resist the popular view that some explanations are non-causal by virtue of being mathematical explanations. To support his objection, Strevens provides a discussion of various explanations that other philosophers regard as instances of non-causal *qua* being mathematical explanations (such as equilibrium explanations and statistical explanations). He argues that, at least in the context of these examples, the mathematical component of an explanation helps scientists to get a better understanding of (or a better grasp on) the relevant causal components cited in the explanation. Hence, Strevens's contribution could be read as defending a limited and careful version of causal reductionism. That is, at least with respect to the examples discussed, there is no reason to question the hegemony of causal accounts.

James Woodward's contribution displays monist tendencies, as he explores whether and to what extent his well-known version of the counterfactual theory of explanation can be extended from its original causal interpretation to certain cases of non-causal explanation. Woodward defends the claim that such an extension is possible in at least two cases: first, if the relevant explanatory counterfactuals do not have an interventionist interpretation, and, second, if the truth of the explanatory counterfactuals is

supported by conceptual and mathematical facts. Finally, he discusses the role of information about irrelevant factors in (non-causal) scientific explanations.

Part II consists of contributions discussing detailed case studies of non-causal explanations from specific scientific disciplines. The case studies under discussion range from neuroscience over earth science to physics. The ambition of these chapters is to analyse in detail what makes a specific kind of explanation from one particular discipline non-causal.

Alisa Bokulich analyses a non-causal explanation from the earth sciences, more specifically from *aeolian geomorphology* (the study of landscapes that are shaped predominantly by the wind). Her case study consists in an explanation of regular patterns in the formation of sand ripples and dunes in deserts of different regions of earth and other planets. Bokulich uses this case study to argue for the “common core conception of non-causal explanation” in order to sharpen the concept of the non-causal character of an explanation. Moreover, she emphasizes that if one has a non-causal explanation for a phenomenon this does not exclude that there is also a causal explanation of the same explanandum.

Mazviita Chirimuuta focuses on a case study from neuroscience, *efficient coding explanation*. According to Chirimuuta, one ought to distinguish four types of explanations in neuroscience: (a) aetiological explanations, (b) mechanistic explanations, (c) non-causal mathematical explanations, and (d) efficient coding explanations. Chirimuuta argues that efficient coding explanations are distinct from the types (a)–(c) and are an often overlooked kind of explanation whose explanatory resources hinge on the implementation of an abstract coding scheme or algorithm. Chirimuuta explores ways in which efficient coding explanations go ‘beyond causation’ in that they differ from mechanistic and, more broadly, causal explanations. The global outlook of Chirimuuta’s chapter is monist in its spirit, as she indicates that all four types of explanations—including efficient coding explanations—answer what-if-things-had-been-different questions which are at the heart of counterfactual theories.

Steven French and **Juha Saatsi** investigate explanations from physics that turn on *symmetries*. They argue that a counterfactual-dependence account, in the spirit of Woodward, naturally accommodates various symmetry explanations, turning on either discrete symmetries (e.g., permutation invariance in quantum physics), or continuous symmetries (supporting the use of Noether’s theorem). The modal terms in which French and Saatsi account for these symmetry explanations throw light on the debate regarding the explanatory status of the Pauli exclusion principle, for example, and opposes recent analyses of explanations involving Noether’s theorem.

Margaret Morrison provides a rigorous analysis of the non-causal character of *renormalization group explanations* of universality in statistical mechanics. Morrison argues that these explanations exemplify *structural explanations*, involving a particular kind of transformation and the determination of ‘fixed points’ of these transformations. Moreover, Morrison discusses how renormalization group explanations exhibit important differences to other statistical explanations in the context of statistical mechanics

that operate by “averaging over microphysical details”. Although Morrison does not address the issue explicitly, it is clear that she rejects causal reductionism, and it is plausible to say that her non-causal characterization of renormalization group explanations is compatible with pluralism and monism.

Part III extends the analysis of non-causal explanations from the natural and social sciences to extra-scientific explanations. More precisely, the contributions in this part discuss explanatory proofs in pure mathematics and grounding explanations in metaphysics.

Mark Colyvan, John Cusbert, and Kelvin McQueen provide a theory of explanatory proofs in pure mathematics (aka intra-mathematical explanations). An explanatory proof does not merely show that a theorem is true but also why it is true. Colyvan, Cusbert, and McQueen pose the question whether explanatory proofs all share some common feature that renders them explanatory. According to their view, there is no single feature that makes proofs explanatory. Rather one finds at least two types of explanation at work in mathematics: constructive proofs (whose explanatory power hinges on dependence relations) and abstract proofs (whose explanatory character consists in their unifying power). Constructive and abstract proofs are two distinct ‘flavours’ of explanation in pure mathematics requiring different philosophical treatment. In other words, Colyvan, Cusbert, and McQueen make the case for explanatory pluralism in the domain of pure mathematics.

Lina Jansson analyses non-causal grounding explanations in metaphysics. In the flourishing literature on grounding, there is large agreement that grounding relations are explanatory and that they are explanatory in a non-causal way. But what makes grounding relations explanatory? According to some recent ‘interventionist’ approaches, the answer to this question should begin by assuming that grounding is a relation that is closely related to causation and, more precisely, that grounding explanations should be given an account in broadly interventionist terms (relying on structural equations and directed graphs functioning as representations of grounding relations). If these interventionist approaches were successful, they would provide a unified monist framework for ordinary causal and grounding explanations. However, Jansson argues that interventionist approaches to grounding explanations fail because causal explanations and grounding explanations differ with respect to the aptness of the causal models and grounding models underlying the explanations.

References

- Achinstein, P. (1983), *The Nature of Explanation* (New York: Oxford University Press).
 Andersen, H. (2014), ‘A Field Guide to Mechanisms: Part I’, *Philosophy Compass* 9: 274–83.
 Bartelborth, T. (1996), *Begründungsstrategien* (Berlin: Akademie Verlag).
 Batterman, R. (2000), ‘Multiple Realizability and Universality’, *British Journal for the Philosophy of Science* 51: 115–45.
 Batterman, R. (2002), *The Devil in the Details* (New York: Oxford University Press).

- Bliss, R. and Trogdon, K. (2016), 'Metaphysical Grounding', *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.). <<https://plato.stanford.edu/archives/win2016/entries/grounding/>>.
- Craver, C. (2007), *Explaining the Brain* (New York: Oxford University Press).
- Craver, C. and Tabery, J. (2017), 'Mechanisms in Science', *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.). <<https://plato.stanford.edu/cgi-bin/encyclopedia/archinfo.cgi?entry=science-mechanisms&archive=spr2017>>.
- Davidson, D. (1980), *Essays on Actions and Events* (Oxford: Oxford University Press).
- Forge, J. (1980), 'The Structure of Physical Explanation', *Philosophy of Science* 47: 203–26.
- Forge, J. (1985), 'Theoretical Explanations in Physical Science', *Erkenntnis* 23: 269–94.
- Friedman, M. (1974), 'Explanation and Scientific Understanding', *Journal of Philosophy* 71: 5–19.
- Frisch, M. (1998), 'Theories, Models, and Explanation', Dissertation, UC Berkeley.
- Hempel, C. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press).
- Hüttemann, A. (2004), *What's Wrong With Microphysicalism?* (London: Routledge).
- Kitcher, P. (1984), *The Nature of Mathematical Knowledge* (Oxford: Oxford University Press).
- Kitcher, P. (1989), 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. Salmon (eds.), *Minnesota Studies in the Philosophy of Science, Vol. 13: Scientific Explanation* (Minneapolis: University of Minnesota Press), 410–505.
- Lange, M. (2016), *Because Without Cause: Non-Causal Explanations in Science and Mathematics* (New York: Oxford University Press).
- Lewis, D. (1986), 'Causal Explanation', in *Philosophical Papers Vol. II* (New York: Oxford University Press), 214–40.
- Lipton, P. (1991/2004), *Inference to the Best Explanation* (London: Routledge).
- Mach, E. (1905), *Erkenntnis und Irrtum. Skizzen zur Psychologie der Forschung* (Leipzig: Barth).
- Mancosu, P. (2015), 'Explanation in Mathematics', *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zalta (ed.). <<https://plato.stanford.edu/archives/sum2015/entries/mathematics-explanation/>>.
- Nerlich, G. (1979), 'What Can Geometry Explain?', *British Journal for the Philosophy of Science* 30: 69–83.
- Price, H. (1996), *Time's Arrow and Archimedes' Point* (Oxford: Oxford University Press).
- Price, H. and Corry, R. (eds.) (2007), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (Oxford: Clarendon Press).
- Reutlinger, A. (2017), 'Explanation Beyond Causation? New Directions in the Philosophy of Scientific Explanation', *Philosophy Compass*, Online First, DOI: 10.1111/phc3.12395.
- Ruben, D.-H. (1990/2012), *Explaining Explanation* (Boulder, CO: Paradigm Publishers).
- Russell, B. (1912/13), 'On the Notion of Cause', *Proceedings of the Aristotelian Society* 13: 1–26.
- Salmon, W. (1989), *Four Decades of Scientific Explanation* (Pittsburgh, PA: University of Pittsburgh Press).
- Scheibe, E. (2007), *Die Philosophie der Physiker* (München: C. H. Beck).
- Skow, B. (2014), 'Are There Non-Causal Explanations (of Particular Events)?', *British Journal for the Philosophy of Science* 65: 445–67.
- Skow, B. (2016), *Reasons Why* (Oxford: Oxford University Press).

- Sober, E. (1983), 'Equilibrium Explanation', *Philosophical Studies* 43: 201–10.
- Strevens, M. (2008), *Depth* (Cambridge, MA: Harvard University Press).
- Van Fraassen, B. (1980), *The Scientific Image* (Oxford: Oxford University Press).
- Van Fraassen, B. (1989), *Laws and Symmetry* (Oxford: Oxford University Press).
- Woodward, J. (2003), *Making Things Happen* (New York: Oxford University Press).
- Woodward, J. (2014), 'Scientific Explanation', *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.). <<https://plato.stanford.edu/archives/win2014/entries/scientific-explanation/>>.
- Wright, G. H. von (1971), *Explanation and Understanding* (Ithaca: Cornell University Press).
- Wright, G. H. von (1974), *Causality and Determinism* (New York and London: Columbia University Press).

PART I

General Approaches

1

Because Without Cause

Scientific Explanations by Constraint

Marc Lange

1. Introduction

Some scientific explanations are not causal explanations in that they do not work by describing contextually relevant features of the world's network of causal relations. Here is a very simple example (inspired by Braine 1972: 144):

Why does Mother fail every time she tries to distribute exactly 23 strawberries evenly among her 3 children without cutting any (strawberries—or children!)? Because 23 cannot be divided evenly into whole numbers by 3.

In a closely related non-causal explanation, the explanandum is simply Mother's failure *on a given occasion* to distribute her strawberries evenly among her children (without cutting any), and the explanans is that Mother has 3 children and 23 strawberries on that occasion and that 23 cannot be divided evenly by 3. Although Mother's having 3 children and 23 strawberries are causes of her failure on this occasion, this explanation does not acquire its explanatory power by virtue of specifying causes. Rather, Mother's strawberries were not distributed evenly among her children because (given the numbers of strawberries and children) they *cannot* be. The particular causal mechanism by which she tried to distribute the strawberries does not enter into it. Even a physically impossible causal mechanism (as long as it is mathematically possible) would have failed.¹

Similar remarks apply to explaining why no one ever succeeded in untying a trefoil knot or in crossing all of the bridges of Königsberg exactly once (while remaining always on land and taking a continuous path)—with the bridges as they were in 1735, when Euler showed that such an arrangement of bridges (let's call it “arrangement K”) *cannot* be crossed. These explanations explain why every attempt to perform a given

¹ Although the explanandum holds with mathematical necessity, this is a scientific explanation rather than an explanation in mathematics: the explanandum concerns a concrete, spatiotemporal system, not exclusively abstract mathematical objects or structures. Everything I say in this chapter should be understood as limited to scientific explanations. (I discuss explanations in mathematics in my 2014 and 2016.)

task failed. These explanations work not by describing the world's causal relations, but rather by revealing that the performance of the task (given certain features understood to be constitutive of that task) is *impossible*, so the explanandum is *necessary*—in particular, more necessary than ordinary causal laws are. The mathematical truths figuring in the above non-causal explanations possess a stronger variety of necessity (“mathematical necessity”) than ordinary causal laws possess.²

Like mathematical truths, some laws of nature have generally been regarded as modally stronger than the force laws and other ordinary causal laws. For example, the Nobel laureate physicist Eugene Wigner (1972: 13) characterizes the conservation laws in classical physics as “transcending” the various particular kinds of forces there happen to be (e.g., electromagnetic, gravitational, etc.). In other words, energy, linear momentum, angular momentum, and so forth would still have been conserved even if there had been different forces instead of (or along with) the actual forces. It is not the case that momentum is conserved because electrical interactions conserve it, gravitational interactions conserve it, and so forth for each of the actual kinds of fundamental interactions. Rather, every actual kind of fundamental interaction conserves momentum for the same reason: that the law of momentum conservation requires it to do so. The conservation law limits the kinds of interactions there could have been, making a non-conservative interaction impossible. This species of impossibility is stronger than ordinary physical impossibility (though weaker than mathematical impossibility).

Accordingly, the conservation laws power non-causal explanations that are similar to the explanation of Mother's failure to distribute her strawberries evenly among her children. Here is an example from the cosmologist Hermann Bondi (1970: 266; 1980: 11–14). Consider a baby carriage with the baby strapped inside so that the baby cannot separate much from the carriage. Suppose that the carriage and baby are initially at rest, the ground fairly smooth and level, and the carriage's brakes disengaged so that there is negligible friction between the ground and the wheels. (The baby's mass is considerably less than the carriage's.) Now suppose that the baby tosses and turns, shaking the carriage in many different directions. Why, despite the baby's pushing back and forth on the carriage for some time, is the carriage very nearly where it began? Bondi gives an explanation that, he says (let's suppose correctly), transcends the details of the various particular forces exerted by the baby on the carriage. Since there are negligible horizontal external forces on the carriage-baby system, the system's horizontal momentum is conserved; it was initially zero, so it must remain zero. Therefore, whatever may occur within the system, its center of mass cannot begin to move horizontally. The only way for the carriage to move, while keeping the system's center of mass stationary, is for the baby to move in the opposite direction. But since the baby is strapped into the carriage, the baby cannot move far without the carriage moving in about the same way. So the carriage cannot move much.

² The literature on distinctively mathematical explanations in science includes Baker (2009); Lange (2013); Mancosu (2008); and Pincock (2007).

The law that a system's momentum in a given direction is conserved, when the system feels no external force in that direction, can supply this "top-down" explanation because this law holds "irrespective of what goes on inside that system" (Bondi 1970: 266). It would still have held even if there had been kinds of forces inside the system other than those covered by the actual force laws. For this reason, Bondi calls momentum conservation a "super-principle", echoing Wigner's remark about its transcending the force laws.³ It constrains the kinds of forces there could have been just as the fact that 23 cannot be divided evenly by 3 constrains the ways Mother could have distributed her strawberries among her children.

Accordingly, I suggest in this chapter that some scientific explanations (which I dub "explanations by constraint") work not by describing the world's causal relations, but rather by describing how the explanandum involves stronger-than-physical necessity by virtue of certain facts ("constraints") that possess some variety of necessity stronger than ordinary causal laws possess. This chapter aims to clarify how explanations by constraint operate.

One obstacle facing a philosophical account of explanations by constraint is that the account cannot make use of the resources that we employ to understand causal explanations. For instance, consider the law that the electric force on any point charge Q exerted by any long, linear charge distribution with uniform charge density λ at a distance r is equal (in Gaussian CGS units) to $2Q\lambda/r$. This "line-charge" law is causally explained by Coulomb's law, since the force consists of the sum of the forces exerted by the line charge's pointlike elements, and the causes of each of these forces are identified by Coulomb's law. Thus, to account for the explanatory priority of Coulomb's law over the line-charge law, we appeal to the role of Coulomb's law in governing the fundamental causal processes at work in every instance of the line-charge law. But the order of explanatory priority in explanations by constraint cannot be accounted for in this way, since explanations by constraint are not causal explanations. For example, the momentum conservation law is explanatorily prior to the "baby-carriage law" ("Any system consisting of...[a baby carriage in the conditions I specified] moves only a little"), where both of these laws have stronger necessity than ordinary causal laws do. But the order of explanatory priority between these two laws cannot be fixed by features of the causal network.

Likewise, consider the fact that the line-charge law's derivation from Coulomb's law loses its explanatory power if Coulomb's law is conjoined with an arbitrary law (e.g., the law giving a pendulum's period as a function of its length). To account for this loss of explanatory power, we appeal to the pendulum law's failure to describe the causal processes operating in instances of the line-charge law. But since explanations by constraint do not work by describing causal processes, we cannot appeal to those processes to account for the fact that the baby-carriage law's derivation from linear momentum

³ Without citing Bondi, Salmon (1998: 73, 359) also presents this example as an explanation that contrasts with the bottom-up explanation citing the particular forces exerted by the baby.

conservation loses its explanatory power if an arbitrary constraint (e.g., energy conservation) joins momentum conservation as a premise.

We also cannot account for this derivation's failure to explain the baby-carriage law on the grounds that the energy-momentum premise is stronger than it needs to be in order to entail the explanandum (since momentum conservation by itself suffices). After all, even momentum conservation (which explains the baby-carriage law) is stronger than it needs to be in order to entail the explanandum. Why, then, is the momentum conservation law explanatorily relevant (despite being broader than it needs to be) but an even broader constraint does not explain? We cannot answer this question in the same way as accounts of causal explanation answer the analogous question in the case of the line-charge law. Although Coulomb's law is broader than it needs to be in order to entail the line-charge law, all instances of Coulomb's law involve the same kind of fundamental causal interaction. But it is not the case that all instances of momentum conservation involve the same kind of fundamental causal interaction. Indeed, an explanation by constraint works precisely by providing information about the way that the explanandum arises from laws spanning diverse kinds of causal interactions. As "constraints", those laws do not depend on the particular kinds of interactions there actually happen to be.

In section 2, I will distinguish three varieties of explanation by constraint (differing in the kind of explanandum they involve). Then I will set out two important historical examples of proposed explanations by constraint against which we will test our ideas about how such explanations work. These examples involve special relativity's explanation of the Lorentz transformations and Hertz's proposed explanation of the inverse-square character of fundamental forces. These examples will allow us to combat the view (entailed by some accounts of scientific explanation, such as Woodward 2003) that "explanations by constraint" are not genuine scientific explanations. In section 3, I will specify the sense in which constraints are modally stronger than ordinary causal laws. I will also introduce a distinction between "explanatorily fundamental" and "derivative" constraints, which is all of the equipment that I will need in section 4 in order to elaborate the way in which explanations by constraint work: roughly, by supplying information about the source of the explanandum's necessity (just as causal explanations work by supplying information about the explanandum's causal history or, more broadly, about the world's causal network). This account will allow us to understand why certain deductions of constraints exclusively from other constraints lack explanatory power. For instance, I will be able to account for the fact that the baby-carriage law's derivation from linear momentum conservation loses its explanatory power if energy conservation is added to the explanans. Finally, in section 5, I will turn to the order of explanatory priority among constraints. I will argue that there is no fully general ground for the distinction between "explanatorily fundamental" and "derivative" constraints. Rather, the order of explanatory priority among constraints is grounded differently in different cases. I will identify how that order is grounded in relativity's explanation of the Lorentz

transformations and—differently—in Hertz’s proposed explanation of the inverse-square character of fundamental forces.

2. Three Varieties of Explanation by Constraint

Three kinds of “explanations by constraint” can be distinguished on the basis of the kind of explanandum they target. In the first kind, the explanandum is a constraint (that is, it has greater modal strength than ordinary laws of nature do). For example, it could be the fact that whenever Mother tries to distribute 23 strawberries evenly among her 3 children (without cutting any), she fails. Because the explanandum is a constraint, the explanans consists entirely of constraints, since the explanandum cannot depend on any facts possessing less necessity than it does.⁴ Using “c” for “constraint”, I will call this a “type-(c)” explanation by constraint.

By contrast, in the second kind of explanation by constraint, the explanandum is *not* a constraint. For example, suppose we explain why it is that whenever Mother tries to distribute her strawberries evenly among her children (without cutting any), she fails. This explanandum does not specify the numbers of children and strawberries, so it is not a constraint. Therefore, the explanans does not consist entirely of constraints; it includes the non-constraints that Mother has 23 strawberries and 3 children. Using “n” (for “not”) to remind us that the explanandum is *not* a constraint, I will call this a “type-(n)” explanation by constraint.

Finally, in the third type of explanation by constraint, the explanandum is a modal fact: that a given fact is a constraint. For example, whereas the explanandum in one type-(c) explanation is the fact that no one has ever managed to cross bridges in arrangement K, we could instead have asked why *it is impossible* to cross bridges in arrangement K, where the relevant species of impossibility is understood to be stronger than the ordinary physical impossibility of, for example, violating Coulomb’s law. The explanans is that certain other facts possess the same (or stronger) species of modality, entailing that the fact figuring in the explanandum does, too. Using “m” (for “modal”) to remind us that the explanandum is a modal fact, I will call this a “type-(m)” explanation by constraint.

The same threefold distinction can be drawn in the baby-carriage example: we might ask “Why does any system consisting of . . . [a baby carriage in the conditions I described earlier] move only a little?” (type (c)), “Why does this baby carriage move

⁴ Of course, the truth of modally weaker laws can entail the truth of modally stronger laws (without explaining why they are true), just as p can entail q even if p is contingent and q possesses some grade of necessity. For example, q can be $(p \text{ or } r)$ where it is a natural law that r —or even a logical truth that r . I am inclined, however, to insist that p cannot explain why $(p \text{ or } r)$ obtains, since presented as an explanation, p misrepresents $(p \text{ or } r)$ ’s modal status. At least, p does not give a *scientific* explanation of $(p \text{ or } r)$. Some philosophers say that p “grounds” $(p \text{ or } r)$, specifying what it is in virtue of which $(p \text{ or } r)$ holds—and that r does likewise—and that such grounding is a kind of explanation. But I do not see p as thereby explaining why $(p \text{ or } r)$ holds. That is not because r also holds; by the same token I do not see p as explaining why $(p \text{ or } \sim p)$ holds. That is not a scientific explanation.

only a little?” (type (n)), or “Why is it impossible (no matter what forces are at work) for any system consisting of... to move more than a little?” (type (m)). This threefold distinction enables us to ask questions about the relations among these various types of explanation. For instance, the same constraint that helps to explain (type (n)) why a given baby carriage moves only a little also helps to explain (type (c)) why any system consisting of a baby carriage in certain conditions moves only a little. Is there some general relation between type-(c) and type-(n) explanations? I shall propose one in section 4.

We might likewise ask about the relation between type-(c) and type-(m) explanations. That it is impossible (whatever forces may be at work) for a system’s momentum in a given direction to change, when the system feels no external force in that direction, explains (type (m)) why it is similarly impossible for any baby-carriage system (of a given kind, under certain conditions) to move much. Now suppose the explanandum is not that it is *impossible* for such a system to move much, but merely that no such system in fact moves much. Having switched from a type-(m) to a type-(c) explanation, does the explanans remain that momentum conservation is a constraint? Or is the explanans merely that momentum is conserved, with no modality included in the explanans—though in order for this explanation to succeed, momentum conservation must be a constraint?⁵ What difference does it make whether momentum conservation’s status as a constraint is included in the explanans or merely required for the explanation to succeed? I will return to this question in section 4.

We might also ask whether certain deductions of constraints exclusively from other constraints lack explanatory power. Consider the question “Why has every attempt to cross bridges in arrangement K while wearing a blue suit met with failure?” Consider the reply “Because it is impossible to cross such an arrangement while wearing a blue suit.” That no one succeeds in crossing that arrangement while wearing a blue suit is a constraint. But of course, it is equally impossible for someone to cross such an arrangement of bridges whatever clothing (if any) he or she may be wearing. So is the reply “Because it is impossible to cross such an arrangement while wearing a blue suit” no explanation or merely misleading? I shall return to this matter in section 4.

To better understand explanations by constraint, it is useful to have in mind some further examples from the history of science. Consider the standard explanation of why the Lorentz transformations hold.⁶ (According to special relativity, the Lorentz

⁵ Compare Hempel’s D-N model: for the expansion of a given gas to be explained by the fact that the gas was heated under constant pressure and that all gases expand when heated under constant pressure, this last regularity must be a law. But the explanans includes “All gases expand when heated...”, not “It is a law that all gases expand when heated...”.

⁶ Brown (2005) has recently departed from this standard explanation by regarding the Lorentz transformations as dynamic rather than kinematic—that is, as depending on features of the particular kinds of forces there are. I agree with Brown that there is a dynamic explanation of the difference in behavior of a given clock or measuring rod when moving as compared to at rest (having to do with the forces at work within it). But unlike Brown, I do not think that the general Lorentz transformations can be explained dynamically. The transformations do not reflect the particular kinds of forces there happen to be. It is no

transformations specify how a pointlike event's space-time coordinates (x', y', z', t') in one inertial reference frame S' relate to its coordinates (x, y, z, t) in another such frame S .) Einstein (1905) originally derived the Lorentz transformations from the "principle of relativity" (that there is a frame S such that for any frame S' in any allowed uniform motion relative to S , the laws in S and S' take the same form) and the "light postulate" (that in S , light's speed is independent of the motion of its source). However, Einstein and others quickly recognized that the light postulate does not help to explain why the Lorentz transformations hold; the transformations do not depend on anything about the particular sorts of things (e.g., electromagnetic fields) that happen to populate spacetime. (In a representative remark, Stachel (1995: 270–2) describes the light postulate as "an unnecessary non-kinematical element" in Einstein's original derivation.) Today the standard explanation of the Lorentz transformations appeals to the principle of relativity, various presuppositions implicit in the very possibility of two such reference frames (such as that all events can be coordinatized in terms of a globally Euclidean geometry), that the functions X and T in the transformations $x' = X(t, v, x, y, z)$ and $t' = T(t, v, x, y, z)$ are differentiable, and that the velocity of S in S' as a function of the velocity of S' in S is continuous and has a connected domain. These premises are all constraints; they all transcend the particular dynamical laws that happen to hold. For example, physicists commonly characterize the principle of relativity as "a sort of 'super law'" (Lévy-Leblond 1976: 271; cf. Wigner 1985: 700) where "all the laws of physics are constrained" by it; likewise, Earman (1989: 155) says that the special theory of relativity "is not a theory in the usual sense but is better regarded as a second-level theory, or a theory of theories that constrains first-level theories". These premises entail that the transformation laws take the form

$$\begin{aligned}x' &= (1 - kv^2)^{-1/2} (x - vt) \\t' &= (1 - kv^2)^{-1/2} (-kvx + t)\end{aligned}$$

for some constant k . The final premise needed to derive the Lorentz transformations is the law that the "spacetime interval" $I = \left([\Delta x]^2 + [\Delta y]^2 + [\Delta z]^2 - c^2 [\Delta t]^2 \right)^{1/2}$ between any two events is invariant (i.e., equal in S and in S') where c is "as yet arbitrary, and need not be identified with the speed of light", as Lee and Kalotas (1975: 436) say in emphasizing that the transformation laws are not owing to the laws about any particular force or other spacetime inhabitant (such as light). Given the forms that the transformations were just shown to have, the interval's invariance entails that

$$k = c^{-2}$$

Thus we arrive at the Lorentz transformations. (Oftentimes instead of the interval's invariance, an explanation cites the existence of a finite invariant speed c . This is a

coincidence that two rods (or two clocks), constructed very differently, behave in the same way when in motion; this phenomenon does not depend on the particular kinds of forces at work. See Lange (2016).

trivial consequence of—and is explained by—the interval’s invariance.⁷) This explanation depicts the Lorentz transformations as arising entirely from constraints—that is, from principles that are modally stronger than the various force laws and so would still have held, regardless of the kinds of forces there were.

Another candidate type-(c) explanation by constraint, which was proposed by Hertz, may not turn out to succeed fully. But an adequate account of scientific explanation must at least leave room for an explanation of the kind Hertz proposed. In his 1884 Kiel lectures, Hertz said that (as far as science has been able to discover) all fundamental forces that are functions of distance are proportional to the inverse-square of the separation—and that this regularity has never been thought coincidental [*zufällig*] (Hertz 1999: 68). By this, Hertz meant that this regularity is not explained by gravity’s being inverse-square, electrostatic forces’ being inverse-square, and so forth for every kind of fundamental force. Rather, fundamental forces are obliged to be inverse-square. It is a constraint; it has a stronger variety of necessity than any of the force laws. Hertz’s proposed explanation appeals to another fact that he takes to constrain any force there might have been: that every fundamental force acts by contact (that is, by a field at the same spacetime point as the acceleration that it causes) rather than by action at a distance. Consider a configuration of bodies and any imaginary surface enclosing them. If a given sort of influence operates by contact action, then the influence of those bodies on any body outside of the surface must pass through the intervening surface. Therefore, any two configurations with the same field at all points on the surface must have the same field everywhere outside of the surface. As Hertz (1999: 68) rightly notes, the existence of such a “uniqueness theorem” rules out a force that declines linearly with distance or with the cube of the distance. Indeed, for a $1/r^n$ force, a uniqueness theorem holds (in three-dimensional space) only for $n = 2$ (Bartlett and Su 1994). That is why (according to Hertz) all of the various fundamental forces are inverse-square forces.

Regarding these two proposed explanations by constraint, we can ask precisely the sorts of questions that we posed in section 1. What makes the principle of relativity and the spacetime interval’s invariance explanatorily prior to the Lorentz transformations? What makes the three-dimensionality of space and the fact that all fundamental forces operate through fields explanatorily prior to the fact that those forces are all inverse-square? Why is it that even if Hertz’s explanation is correct, the derivation of the inverse-square character of all fundamental forces from the contact-action constraint loses its explanatory power if an arbitrary constraint (e.g., the spacetime interval’s invariance) is added as a premise?

⁷ It is indeed trivial. Suppose that in frame S , a process moving at speed c links two events. Since distance is speed times time, $([\Delta x]^2 + [\Delta y]^2 + [\Delta z]^2 = c^2 \Delta t^2)$, and so the interval I between these events is 0. By I ’s invariance, the two events are separated by $I = 0$ in S' , so $([\Delta x']^2 + [\Delta y']^2 + [\Delta z']^2 - c^2 [\Delta t']^2)^{1/2} = 0$, so the speed in S' of the process linking these events is $([\Delta x']^2 + [\Delta y']^2 + [\Delta z']^2)^{1/2} / \Delta t' = c$. Hence the speed c is invariant. For examples of this standard explanation of the Lorentz transformations, see any number of places; for especially careful discussions, see Aharoni (1965: 12–14); Berzi and Gorini (1969); and Lévi-Leblond (1976).

Some of the arguments that I have termed “explanations by constraint” are deemed to be explanatorily impotent by some accounts of scientific explanation. For instance, according to Woodward (2003), an explanans must provide information about how the explanandum would have been different under various counterfactual changes to the variables figuring in the explanans:

[I]t is built into the manipulationist account of explanation I have been defending that explanatory relationships must be change-relating: they must tell us how changes in some quantity or magnitude would change under changes in some other quantity. Thus, if there are generalizations that are laws but that are not change-relating, they cannot figure in explanations.

(Woodward 2003: 208)

[I]f some putative explanandum cannot be changed or if some putative explanans for the explanandum does not invoke variables, changes in which would be associated with changes in the explanandum, then we cannot use that explanans to explain the explanandum.

(Woodward 2003: 233)

These criteria fail to accommodate some explanations by constraint. Unlike the charges and distances in Coulomb’s law, there are no obvious variables to be changed in the principle of relativity or in the law that every fundamental force acts by contact. Of course, we could insist on treating “action by contact” as the value of a variable in the law that all fundamental forces act by contact, and we might then ask what force laws would have been like had that variable’s value instead been “action at a distance”. But the answer is: any force law might have held. The argument from action by contact to the law that all fundamental forces must be inverse-square reveals nothing about how forces would have varied with separation had they operated by action at a distance; it does not follow, for example, that all fundamental forces would then have been inverse-cube. The argument simply goes nowhere if “action by contact” is changed to “action at a distance”, since under action at a distance, a force does not have to satisfy a uniqueness theorem.⁸

The same goes for changing the principle of relativity in the explanation of the Lorentz transformations. However, if we replace the spacetime interval’s invariance with the invariance of temporal intervals, then the argument *does* yield an alternative to the Lorentz transformations: the Galilean transformations. Indeed, this is the standard explanation given in classical physics of why the Galilean transformations hold. But it is difficult to know whether the replacement of *I*’s invariance with *t*’s invariance should count by Woodward’s lights as a change in the value of a variable in a law, rather than the wholesale replacement of one law with another.

⁸ Woodward (2003: 220–1) compares his own account of causal explanations to Steiner’s (1978a, 1978b) account of explanations in mathematics. But one problem with Steiner’s approach is that when some explanatory proofs are deformed to fit a different class in what is presumably the same “family”, the proofs simply go nowhere rather than yielding a parallel theorem regarding that other class (see Lange 2014). Thus, it is not always the case that “in an explanatory proof we see how the theorem changes in response to variations in other assumptions” (Woodward 2003: 220).

The failure of Woodward's account to allow for typical explanations by constraint is not very surprising. As we just saw, Woodward says that a putative explanandum must be capable of being changed, if certain other conditions change. In contrast, the explanandum in a type-(c) explanation by constraint is a constraint—a fact having an especially strong resistance to being changed. Woodward takes the value of Newton's gravitational constant G as having no explanation in classical gravitational theory because:

[f]rom the point of view of Newtonian gravitational theory, G is a constant, which cannot be changed by changing other variables. . . . To explain something, we must be able to think of it as (representable by) a *variable*, not as a constant or fixed parameter. (Woodward 2003: 234)

What Woodward says about Newton's G applies even more strongly to constraints. Although Woodward allows for non-causal explanations, he insists that both causal and non-causal explanations "must answer what-if-things-had-been-different questions" (Woodward 2003: 221). But consider an explanation by constraint such as "Every kind of force at work in this spacetime region conserves momentum because a force that fails to conserve momentum is impossible; momentum conservation constrains the kinds of forces there could have been." This explanation reveals nothing about the kinds of forces there would have been, had momentum conservation not been a constraint. Like G 's value, momentum conservation is "fixed" in classical physics. However, this explanation does reveal that even if there had been different kinds of forces, momentum would still have been conserved. In this example, information about the conditions under which the explanandum would have remained the same seems to me just as explanatorily relevant as information in Woodward's causal explanations about the conditions under which the explanandum would have been different.

To do justice to scientific practice, an account of scientific explanation should leave room for explanation by constraint. A proposed explanation like Hertz's should be disconfirmed (or confirmed) by empirical scientific investigation, rather than being ruled out a priori by an account of what scientific explanations are.

3. Varieties of Necessity

The idea that I will elaborate is that an explanation by constraint derives its power to explain by virtue of providing information about where the explanandum's especially strong necessity comes from, just as a causal explanation works by supplying information about the explanandum's causal history or the world's network of causal relations. (The context in which the why question is asked may influence what information about the origin of the explanandum's especially strong necessity is relevant; context plays a similar role in connection with causal explanations: by influencing what information about the explanandum's causal history or the world's network of causal relations is relevant.) For instance, the explanandum in a type-(c) explanation by constraint has a stronger variety of necessity than ordinary causal laws (such as force laws) do. A type-(c)

explanation by constraint works, I propose, by supplying some information about the strong kind of necessity possessed by the explanandum and how the explanandum comes to possess it. The explanans may simply be that the explanandum possesses some particular sort of necessity, as in: “Why has no one ever untied a trefoil knot? Not from lack of imagination or persistence, but because it is mathematically impossible to do so.” In many explanations by constraint, however, the explanans does not merely characterize the explanandum as a constraint. Rather, the explanans supplies further information about where the explanandum’s necessity comes from.

For example, an explanation of the “baby-carriage law” may go beyond pointing out that the explanandum transcends the various laws for the particular forces at work in the baby-carriage system. The explanans may also show how the explanandum follows from the law that a system’s horizontal momentum is conserved if the system feels no external horizontal forces, where this law also transcends the various force laws. The explanans thereby supplies considerable information about where the inevitability of the baby-carriage law comes from. It could supply even more information by pointing out that there is nothing special about the horizontal direction; the law about horizontal forces and momentum is necessary because the same law holds of any direction. This constraint, in turn, derives its necessity from that of two others. The first is the fundamental dynamical law relating force to motion (in classical physics: the Euler-Lagrange equation), which possesses exactly the same necessity as the baby-carriage law since the relation between motion and any kind of force also transcends the particular kinds of forces there happen to be. The second is the constraint that if the fundamental dynamical law holds, then linear momentum is conserved. That constraint possesses greater necessity than the fundamental dynamical law since it, in turn, follows from a symmetry principle: that every law is invariant under arbitrary spatial translation. This symmetry principle lies alongside the principle of relativity as a law about laws.

Each of these various, increasingly informative type-(c) explanations of the baby-carriage law supplies information about how the explanandum acquires its especially strong inevitability. Here is a natural way to unpack this idea. The various grades of necessity belong to a pyramidal hierarchy (see Figure 1.1): from strongest at the top to weakest at the bottom. Each rung on the hierarchy consists exclusively of truths possessing the same particular variety of necessity, where none of these truths concerns any truth’s modal status (i.e., its place on or absence from the hierarchy). I shall call these “first-order” truths (and a “first-order claim” is a claim that, if true, states a first-order truth). For example, it is a first-order truth that the momentum of any closed system is conserved. In contrast, a truth not appearing anywhere in this hierarchy (in particular, a “second-order” truth) is that momentum conservation *is a constraint*. Every truth on a given rung is automatically included on the rung immediately below (and so on every rung below), since a truth possessing a given variety of necessity also possesses any weaker variety. (For instance, a mathematical necessity is “by courtesy” physically necessary.) But a given rung also includes some truths absent from the rung

The strongest necessities, including the logical and mathematical truths.

The above together with that (i) momentum is conserved if the Euler-Lagrange equation holds, (ii) the Lorentz transformations hold if the spacetime interval is invariant, (iii) the Galilean transformations hold if the temporal interval is invariant, and others.

The above together with the spacetime interval's invariance, the Lorentz transformations, and others.

The above together with the Euler-Lagrange equation, the conservation laws, the "baby-carriage law," and others.

The above together with the force laws and others.

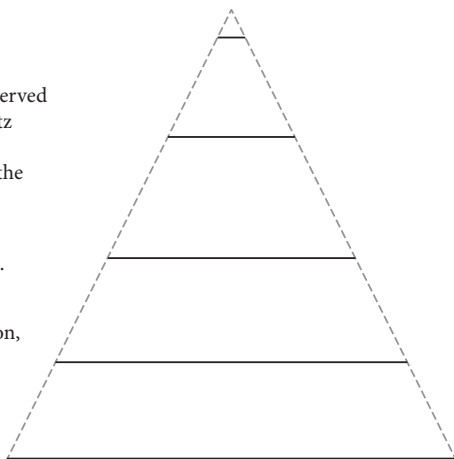


Figure 1.1 Some grades of necessity.

immediately above (and so absent from any rung above). The top rung contains the truths possessing the strongest necessity, including the logical and mathematical truths.⁹ The force laws lie on the bottom rung. Between are various other rungs; the constraints are located somewhere above the bottom rung. For example, the conservation laws do not occupy the highest rung, but since they are constraints, they sit on some rung above the lowest (and on every rung below the highest on which they lie). Every rung is logically closed (in first-order truths), since a logical consequence of a given truth possesses any variety of necessity that the given truth possesses.

If the highest rung on which p appears is higher than the highest rung on which q appears, then p 's necessity is stronger than q 's. This difference is associated with a difference between the ranges of counterfactual antecedents under which p and q would still have held. For instance, a conservation law p , as a constraint on the force laws q , would still have held even if there had been different force laws. Although nothing I say here will turn on this point, I have argued elsewhere (Lange 2009) that the truths on a given rung would all still have held had r obtained, for any first-order claim r that is logically consistent with the truths on the given rung taken together. This entails (I have shown) that the various kinds of necessities must form such a pyramidal hierarchy. In addition to this hierarchy of first-order truths, a similar hierarchy is formed by the varieties of necessity possessed by second-order truths (together with any first-order truths they may entail). For instance, the principle of relativity (that any law takes the same form in any reference frame in a certain family) is a second-order truth (since it says something about the *laws*, i.e., the truths on the bottom rung of the first-order hierarchy), and it is a constraint since it does not lie on the lowest rung of the second-order hierarchy; it does not say simply that a given first-order truth is necessary.

⁹ Perhaps the narrowly logical truths occupy a rung above the mathematical truths. In either case, the mathematical truths transcend the various rungs of natural laws.

On my view, the truths on a given rung of the second-order hierarchy would still have held had r been the case, for any second-order or first-order claim r that is logically consistent with the truths on the given rung. Once again, though nothing I say here will turn on this point, I have shown that if some second-order and first-order truths form a rung on the second-order hierarchy, then the first-order truths on that rung themselves form a rung on the first-order hierarchy.

One way for an explanation by constraint to work is simply by telling us that the explanandum possesses a particular kind of inevitability (strong enough to make it a constraint)—that is, by locating it on the highest rung to which it belongs (somewhere above the hierarchy's lowest rung). But as we have seen, an explanation by constraint can also tell us about how the explanandum comes to be inevitable. To elaborate this idea, we need only to add a bit more structure to our pyramidal hierarchy. A given constraint can be explained only by constraints at least as strong; a constraint's necessity cannot arise from any facts that lack its necessity (see Lange 2008). But a constraint cannot be explained entirely by constraints possessing stronger necessity than it possesses, since then it would follow logically from those constraints and so itself possess that stronger necessity. Accordingly, on a given rung of constraints (i.e., above the hierarchy's lowest rung), there are three mutually exclusive, collectively exhaustive classes of truths:

- First, there are truths that also lie on the next higher rung—truths possessing some stronger necessity.
- Second, there are truths that are not on the next higher rung and that some other truths on the given rung help to explain. Let's call these "explanatorily derivative" laws (or "EDLs" on that rung).
- Third, there are truths that are not on the next higher rung and that no other truths on the given rung help to explain. Let's call these truths the rung's "explanatorily fundamental" laws ("EFLs" on that rung).

I suggest that every EDL on a given rung follows logically from that rung's EFLs together (perhaps) with truths possessing stronger necessity.¹⁰ A type-(c) explanation by constraint explains a given constraint either by simply identifying it as a constraint of a certain kind or by also supplying some information about how its necessity derives from that of certain EFLs. Any EDL can be explained entirely by some EFLs that together entail it: some on its own rung, and perhaps also some on higher rungs.

I have said that when the "baby-carriage law" is given an explanation by constraint, then it is explained by the fact that it transcends the various force laws, and this explanation can be enriched by further information about how its necessity derives

¹⁰ The EFLs on a given rung may be stronger than the minimum needed to supplement the necessities on a higher rung in order to entail all of the EDLs on the given rung. For instance, a proper subset of the EFLs may suffice (together with the stronger necessities) to entail not only all of the EDLs, but also the remaining EFLs. But not all entailments are explanations (of course). Some of the EFLs may entail the others without explaining them. Likewise, perhaps a given EDL could be explained by any of several combinations of EFLs. Of course, a textbook writer might choose as a matter of convenience to regard some of the EFLs as axioms and others as theorems. But that choice would be made on pedagogic grounds; the "axioms" among the EFLs would still not be explanatorily prior to all of the "theorems".

from that of EFLs. I have thereby suggested that the explanans in a type-(c) explanation is not simply some constraint's truth, but the fact that it is a constraint, since the explanation works by supplying information about where the explanandum's necessity comes from. The explanans in a type-(c) explanation thus takes the same form as the explanans in a type-(m) explanation. These are my answers to some of the questions that I asked earlier. We will see another argument for these answers at the end of section 4. (When I say, then, that a given EFL helps to explain a given EDL, I mean that the EFL's *necessity* helps to explain the EDL.)

No truth on a given EFL's own rung—and, therefore, no truth on any higher rung of the hierarchy (since any truth on a higher rung is also on every rung below)—helps to explain that EFL. A truth in the given pyramidal hierarchy that is *not* on the rung for which a given truth is an EFL also cannot help to explain the EFL, since the EFL cannot depend on truths that lack its necessity. An EFL on some rung of the first-order hierarchy may be brute—that is, have no explanation (other than that it holds with a certain kind of necessity). This may be the case, for example, with the fundamental dynamical law (classically, the Euler-Lagrange equation). But an EFL on some rung of the first-order hierarchy may not be brute, but instead be explained by one or more second-order truths (leaving aside the second-order truth that the given EFL is necessary). For example, the constraint that momentum is conserved if the Euler-Lagrange equation holds (which, as I mentioned a moment ago, figures in the explanation of momentum conservation) may have no explanation among first-order truths, but is explained by a second-order truth (namely, the symmetry principle that every law is invariant under arbitrary spatial translation). It is entailed by the symmetry principle, so although it may be an EFL on some rung of the first-order pyramid, it is an EDL on the same rung of the second-order pyramid as the symmetry principle. The same relation holds between the principle of relativity and the constraint that the Lorentz transformations hold if spacetime intervals are invariant (as well as the constraint that the Galilean transformations hold if temporal intervals are invariant). This constraint, together with the spacetime interval's invariance (which may be an EFL), explains why the Lorentz transformations hold (as we saw in section 2).

In section 4, I will argue that this picture allows us to understand why certain deductions of constraints exclusively from other constraints do not qualify as explanations by constraint, thereby addressing some of the questions about explanation by constraint that I posed earlier. Obviously, this picture presupposes a distinction between EFLs and EDLs on a given rung of the hierarchy. In section 5, I will consider what makes a constraint “explanatorily fundamental”.

4. How do Explanations by Constraint Work?

Although any EDL can be explained by being deduced from EFLs on its own rung (together, perhaps, with some on higher rungs), not every such deduction is an explanation. For example, the baby-carriage law is explained by the EFLs responsible

for momentum conservation, but this argument loses its explanatory power (while retaining its validity) if its premises are supplemented with an arbitrary EFL possessing the explanandum's necessity (such as the spacetime interval's invariance). The added EFL keeps the deduction from correctly specifying the EFLs from which the explanandum acquires its inevitability. Accordingly, I propose:

If d (an EDL) is logically entailed by the conjunction of $f, g, \dots$ (each conjunct an EFL on or above the highest rung on which d resides), but g (a logically contingent truth¹¹) is dispensable (in that d is logically entailed by the conjunction of the other premises), then the argument from $f, g, \dots$ does not explain d .

Of course, g may be dispensable to one such argument for d without being dispensable to every other.¹² But (I suggest) if g is dispensable to every such argument, then g is “explanatorily irrelevant” to d —that is, g is a premise in no explanation by constraint of d . In other words, if no other EFLs on d 's rung (or above) combine with g to entail d where g is indispensable to the argument, then no EDLs on d 's rung (or above) render g explanatorily relevant to d . Any power that g may have to join with other constraints to explain d derives ultimately from its power to join with some other EFLs (or its power standing alone) to explain d . This idea is part of the picture (sketched in section 3) of explanations by constraint as working by virtue of supplying information about how the explanandum's necessity derives from the necessity of some EFLs.¹³

If d is an EDL and g is an EFL on a given rung, then even if there are no deductions of d exclusively from EFLs (on or above that rung) to which g is indispensable, there are deductions of d from EDLs and EFLs on d 's rung to which g is indispensable. For example, g is indispensable to d 's deduction from g and $g \supset d$. But g 's indispensability to such a deduction is insufficient to render g explanatorily relevant to d . To be explanatorily relevant to d , an EFL must be indispensable to a deduction of d from EFLs alone. If every logically contingent premise is indispensable to such an argument, then the argument qualifies (I suggest) as an explanation by constraint (type-(c)):

If d (an EDL) is logically entailed by the conjunction of $f, g, \dots$ (each conjunct an EFL on or above the highest rung on which d resides) and the conjunction of no proper subset of $\{f, g, \dots\}$ logically entails d , then the argument explains d .

If g (a logically contingent EFL on or above d 's highest rung) is explanatorily irrelevant to d (an EDL), then in particular, g figures in no explanation of d exclusively from EFLs.

¹¹ By “logically contingent” truths, I mean all but the *narrowly* logical truths. A mathematical truth then qualifies as “logically contingent” because its truth is not ensured by its logical form alone. All and only narrowly logical truths can be omitted from any valid argument's premises without loss of validity.

¹² Even if g is dispensable to one such argument, g may nevertheless entail d . In that case, d would have two explanations by constraint exclusively from EFLs.

¹³ This paragraph addresses Pincock's (2015: 875) worry that I am “working with the idea that an explanation need only cite some sufficient conditions for the phenomenon being explained... [T]here is a risk that redundant conditions will be included. These conditions will not undermine the modal strength of the entailment, so it is not clear why Lange would say they undermine the goodness of the explanation.”

So for any deduction of d exclusively from g and other EFLs on or above d 's highest rung, some logically contingent premise must be dispensable (else that argument would explain d , contrary to g 's explanatory irrelevance to d). If g is not the sole dispensable premise, then suppose one of the other ones is omitted. The resulting argument must still have a dispensable premise, since otherwise it would explain d and so g would be explanatorily relevant to d . If there remain other dispensable premises besides g , suppose again that one of the others is omitted, and so on. Any argument that is the final result of this procedure must have g as its sole dispensable premise—in which case g must have been dispensable originally. Therefore, if g is explanatorily irrelevant to d , then g is dispensable to every deduction of d exclusively from EFLs on or above d 's highest rung. (This is the converse of an earlier claim.)

I began this section by suggesting that an EDL fails to be explained by its deduction exclusively from EFLs on or above its highest rung if one of the deduction's logically contingent premises is dispensable. The distinction between EFLs and EDLs is crucial here; an EDL's deduction from EDLs on its own rung may be explanatory even if some of the deduction's logically contingent premises are dispensable. For example, the baby-carriage law is explained by the law that a system's horizontal momentum is conserved if the system feels no horizontal external forces. Validity does not require the additional premise that the same conservation law applies to any non-horizontal direction. But the addition of this premise would not spoil the explanation. Rather, it would supply additional information regarding the source of the baby-carriage law's inevitability: that it arises from EFLs that in this regard treat all directions alike. The baby-carriage law is explained by the EDL that for any direction, a system's momentum in that direction is conserved if the system feels no external forces in that direction.

An EDL figures in an explanation by constraint in virtue of supplying information about the EFLs that explain the explanandum. It supplies this information because some of those EFLs explain it. Hence, d (an EDL) helps to explain e (another EDL) only if any EFL that helps to explain d also helps to explain e . For example, the spacetime interval's invariance does not help to explain the baby-carriage law, so the Lorentz transformations must not help to explain the baby-carriage law (because the interval's invariance helps to explain the Lorentz transformations).

If we remove the restriction to EFLs, then this idea becomes the transitivity of explanation by constraint: if c helps to explain d and d helps to explain e , then c helps to explain e . Although the literature contains several kinds of putative examples where *causal* relations are intransitive, none of those examples suggests that explanations by constraint can be intransitive. For example (see Lewis 2007: 480–2), event c (the throwing of a spear) causes event d (the target's ducking), which causes event e (the target's surviving), but according to some philosophers, c does not cause e because c initiates a causal process that threatens to bring about $\sim e$ (though is prevented from doing so by d). Whether or not this kind of example shows that causal relations can be intransitive, it has no analogue among explanations by constraint, since they do not reflect causal processes such as threats and preventers. In other putative examples of intransitive

causal relations (see Lewis 2007: 481–2), c (a switch's being thrown) causes d (along some causal pathway), which causes outcome e , but according to some philosophers, c does not cause e if e would have happened (though in a different way) even if $\sim c$. Again, regardless of whether this kind of example demonstrates that token causal relations can be intransitive, explanations by constraint cannot reproduce this phenomenon since they do not aim to describe causal pathways. They involve no switches; if constraint d follows from one EFL on d 's highest rung and follows separately from another, then each EFL suffices to explain d by constraint.¹⁴

I have just been discussing explanations by constraint where the explanandum is a constraint. Earlier I termed these “type-(c)” explanations by constraint. In contrast, a “type-(n)” explanation gives the reason why Mother fails whenever she tries to distribute her strawberries evenly among her children. That reason involves not only constraints, but also the non-constraint that Mother has exactly 23 strawberries and 3 children. This explanation works by supplying information about how Mother's failure at her task, given non-constraints understood to be constitutive of that task, comes to possess an especially strong variety of inevitability.

What about Mother's failure to distribute her strawberries evenly among her children *while wearing a blue suit*? Although that task consists partly of wearing a blue suit, Mother's failure has nothing to do with her attire. Her suit's explanatory irrelevance can be captured by this principle:

Suppose that s and w are non-constraints specifying that the kind of task (or, more broadly, kind of event) in question has certain features. Let w be strictly weaker than s . Suppose that s and some EFLs logically entail that any attempt to perform the task fails (or, more broadly, that no event of the given kind ever occurs), and this failure is not entailed by s and any proper subset of these EFLs. But suppose that w suffices with exactly the same EFLs to logically entail that any attempt fails (or that no such event occurs). Then the argument from s and these EFLs (or EDLs that they entail) fails to explain by constraint why any such attempt fails (or why no such event occurs).

Roughly, if s is stronger than it needs to be, then it includes explanatorily superfluous content. That Mother is wearing a blue suit thus figures in no type-(n) explanations of her failure at her task.

This above principle says that for w to make s “stronger than it needs to be”, w must be able to make do with exactly the same EFLs as s . But a non-constraint s can explain even if it can be weakened without rendering the argument invalid—as long as that weakening must be balanced by the argument's EFLs being strengthened. Let's look at an example. The fact that Mother has exactly 23 strawberries and 3 children is not stronger than it needs to be to entail the explanandum when the other premise

¹⁴ In addition, explanation may sometimes be intransitive because although c explains and entails d , and d explains e , d does not suffice to entail e . Rather, e follows from d only when d is supplemented by premises supplied by the context put in place by the mention of d . In that case, c may neither entail nor explain e (Owens 1992: 16). But explanations by constraint are all deductively valid.

(let's suppose it to be an EFL) is that 3 fails to divide 23 evenly into whole numbers. However, it is stronger than it needs to be to entail the explanandum when the other premise is that 3 fails to divide 23 evenly *and* 2 fails to divide 23 evenly. With this stronger pair of EFLs, the non-constraint premise *s* can be weakened to the fact *w* that Mother has exactly 23 strawberries and 2 or 3 children. Nevertheless, the original, stronger non-constraint is explanatory. Notice that the EFL that 2 fails to divide 23 evenly is not a premise in the original deduction—and had it been, then it would have been dispensable there. Accordingly, the above principle specifying when *s* is stronger than it needs to be requires that the argument from *w* use exactly the same EFLs as the argument from *s* and that each of those EFLs be indispensable to the argument from *s*.¹⁵ Hence, that Mother's task involves her having 23 strawberries and 3 children helps to explain why Mother always fails in her task; this fact about her task requires no weakening to eliminate explanatorily superfluous content, unlike any fact entailing that the task involves Mother's wearing a blue suit.

Any constraint that joins with Mother's having 23 strawberries and 3 children to explain (type-(n)) why Mother fails to distribute her strawberries evenly among her children also explains (type-(c)) why it is that if Mother has 23 strawberries and 3 children, then she fails to distribute her strawberries evenly among her children. Here is a way to capture this connection between type-(c) and type-(n) explanations by constraint:

If there is a type-(n) explanation by constraint whereby non-constraint *n* and constraint *c* explain why events of kind *e* never occur, then there is a type-(c) explanation by constraint whereby *c* explains why it is that whenever *n* holds, *e*-events never occur.

The converse fails, as when *c* is that 3 fails to divide 23 evenly, *n* is that Mother's task involves her having 23 strawberries and 3 children and wearing a blue suit, and *e* is Mother's succeeding at distributing her strawberries evenly among her children; with regard to explaining why *e*-events never occur, *n* contains explanatorily superfluous content.

Suppose that constraint *c* explains (type-(c)) why all attempts to cross bridges in a certain arrangement K while wearing a blue suit fail. Why, then, do all attempts to cross *Königsberg's* bridges while wearing a blue suit fail? This explanandum is not a constraint. Accordingly, the explanans consists not only of *c*, but also of the fact that *Königsberg's* bridges are in arrangement K. But although the explanans in this type-(n) explanation includes that the task involves crossing bridges in arrangement K, it does not include that the task involves doing so while wearing a blue suit; any such content would be explanatorily superfluous. So the same explanans explains why no one ever

¹⁵ Of course, there is a constraint that entails the explanandum when the other premise is that the task involves Mother's having 23 strawberries and 3 children and wearing a blue suit, and where the argument is rendered invalid if the same constraint is used but the other premise is weakened so as not to entail wearing a blue suit. But that constraint is an EDL, not an EFL as the criterion mandates. Thus, the criterion does not thereby render Mother's attire explanatorily relevant.

succeeds in crossing Königsberg's bridges, blue suit or no. Since c and the fact that Königsberg's bridges are in arrangement K explains (type-(n)) why no one ever succeeds in crossing Königsberg's bridges, the above connection between type-(c) and type-(n) explanations entails that c explains (type-(c)) why it is that if Königsberg's bridges are in arrangement K, no one succeeds in crossing them. Presumably, the same applies to bridges anywhere else.

I have just argued that if a constraint explains why all attempts to cross bridges in arrangement K while wearing a blue suit fail, then the same constraint also explains why all attempts to cross bridges in arrangement K fail. By the same kind of argument, any constraint that explains why all *past* attempts to untie trefoil knots failed also explains why all attempts to untie trefoil knots fail. There is no special reason why all past attempts fail.

It might be objected that the fact that every attempt to untie trefoil knots fails obviously does not explain itself but nevertheless explains (by constraint) why, in particular, every past attempt failed. But I do not agree that the fact that every attempt to untie trefoil knots fails explains (by constraint) why every past attempt failed. Rather, the fact that every attempt to untie trefoil knots *must* fail (as a matter of mathematical necessity) explains by constraint why every past attempt failed and likewise why every attempt fails. The explanans in a type-(c) explanation is not simply some constraint's truth, but the fact that it is a constraint. The explanans in a type-(c) explanation thus takes the same form as the explanans in a type-(m) explanation.

5. What Makes a Constraint Explanatorily Fundamental?

The approach I have just sketched depends upon a distinction between EFLs and EDLs among the truths having the same rung (above the lowest) as their highest on the first-order pyramidal hierarchy of necessities. What grounds this distinction among constraints that possess the same variety of necessity?

For example, the standard explanation of the Lorentz transformations (described in section 2) appeals to the principle of relativity to explain why the coordinate transformations must be either the Galilean or the Lorentz transformations, and then appeals to the spacetime interval's invariance to explain why the Lorentz rather than the Galilean transformations hold. But what makes the interval's invariance an EFL? To distinguish the Lorentz from the Galilean transformations in the derivation's final step, any kinematic consequence of special relativity that departs from classical mechanics would suffice, such as the relativistic formula for adding parallel velocities¹⁶ or the relativity of simultaneity. Why is the interval's invariance rather than either of these

¹⁶ In special relativity, the sum of parallel velocities v_1 and v_2 is $(v_1 + v_2)/(1 + v_1 v_2/c^2)$, whereas in classical physics it is $v_1 + v_2$.

an EFL?¹⁷ For that matter, why don't the Lorentz transformations themselves qualify as EFLs and so explain the interval's invariance (which they entail), rather than the reverse? What makes the interval's invariance explanatorily prior to the Lorentz transformations (rather than the reverse, for instance—or the relativity of simultaneity being explanatorily prior to each)?

I believe that there is no fully general reason why certain constraints rather than others on a given rung (but none higher) constitute EFLs. The order of explanatory priority is grounded differently in different cases. A principle sufficiently general to apply to any rung of the hierarchy, no matter what its content, and purporting to specify which constraints are "axioms" (EFLs) and which are "theorems" (EDLs) will find it very difficult to discriminate as scientific practice does between the Lorentz transformations, the interval's invariance, the velocity-addition law, and the relativity of simultaneity. EFLs are set apart from EDLs on specific grounds that differ in different cases rather than on some uniform, wholesale basis.

As an example of how an attractive wholesale approach founders, consider Watkins's (1984: 204–10) criteria for distinguishing "natural" from "unnatural" axiomatizations having exactly the same deductive consequences. He contends that a natural axiomatization contains as (finitely) many axioms as possible provided that

1. each axiom in the axiom set is logically independent of the conjunction of the others
2. no predicate or individual constant occurs inessentially in the axiom set
3. if axioms containing only non-observational predicates can be separately stated, without violating any other rules, then they are separate, and
4. no axiom contains a (proper) component that is a theorem of the axiom set (or becomes one when its variables are bound by the quantifiers that bind them in the axiom).¹⁸

These criteria deem certain axiomatizations to be unnatural. Rule 2, for example, ensures that a natural axiomatization not have as one axiom "A system's horizontal momentum is conserved if the system feels no horizontal external forces" and an analogous constraint for non-horizontal momentum as another, separate axiom. However, Watkins's criteria cannot privilege the interval's invariance over the velocity-addition law, the relativity of simultaneity, or the Lorentz transformations. I see no way for wholesale rules like Watkins's to pick out which of these is an EFL.

¹⁷ I am not asking about the explanatory priority of the principle of relativity because it is not modally on a par with the interval's invariance and the transformation laws; it is not on the same rung as they. Rather, it is a meta-law, belonging to the hierarchy of second-order truths. See Lange (2009).

¹⁸ Watkins intends these criteria for a "natural axiomatization" to determine what counts as a "unified scientific theory" (rather than a "rag-bag 'theory'"); Watkins thereby uses these criteria to elaborate the idea that more fundamental explanations involve more unified theories. Salmon (1998: 401) also tentatively suggests that Watkins's criteria be used to understand scientific explanation.

What, then, grounds the order of explanatory priority among the Lorentz transformations and the other constraints on a modal par with it? What is the main difference between the interval's invariance (and the invariance of some finite speed c , which is explained by following from the interval's invariance and, in turn, explains the Lorentz transformations), on the one hand, and the relativity of simultaneity, the Lorentz transformations, and the velocity-addition law, on the other hand? I suggest that the main difference between them is that the former identifies certain quantities as invariant whereas each of the latter relates frame-dependent features in two frames or within a given frame. The behavior of invariant quantities is explanatorily prior to the behavior of frame-dependent quantities because invariant quantities are features of the world, uncontaminated by the reference frame from which the world is being described, whereas frame-dependent quantities reflect not only the world, but also the chosen reference frame. How things *are* explains how they *appear* from a given vantage point. This view is often expressed by physicists and philosophers alike (Brading and Castellani 2003: 15; Eddington 1920: 181; Mermin 2009: 79; North 2009: 63, 67; Salmon 1998: 259). Reality explains mere appearances, and so the law that a certain quantity is invariant takes explanatory priority over the law specifying how a certain frame-dependent quantity transforms. For the same reason, the Galilean spatial transformations are not treated as EFLs in classical physics; explanations of why they hold (according to classical physics) finish by appealing not to (e.g.) the classical velocity-addition formula, but rather to the law that temporal intervals are invariant (i.e., $\Delta t = \Delta t'$). Time's absolute character is "fundamental" in Newtonian physics (cf. Barton 1999: 12).

But although reality's explanatory priority over appearances grounds the EFL/EDL distinction in this case, it cannot do so generally. In other cases, the distinction must be grounded in other ways. Consider, for example, Hertz's proposed explanation of the fact that all fundamental forces are inverse-square. According to Hertz, what makes the three-dimensionality of space and the fact that all fundamental forces operate through fields explanatorily prior to the fact that those forces are all inverse-square?¹⁹ That reality explains mere appearances cannot account for the order of explanatory priority in this case.

I suggest that the distinction between EFLs and EDLs in this case arises instead from the common idea that features of the spatiotemporal theater are explanatorily prior to features of the actors who strut across that stage. For instance, if it were a law that space has a certain finite volume V , then the fact that no material object's volume exceeds V would be an EDL that is explained by a feature of space: only entities of a certain maximum size could fit within the theater. Space's three-dimensionality is likewise

¹⁹ Hertz's purported explanation also appeals to the existence of "uniqueness theorems" for certain functions but not others. These are mathematical facts, so they occupy a higher rung on the hierarchy than the explanandum. Their explanatory priority is thereby secured.

prior to the features of any of space's denizens, including forces.²⁰ Whereas the fact that all forces are inverse-square concerns a feature of space's occupants, the fact that all forces act by fields rather than at a distance is (for Hertz) more fundamental than that. Hertz sees it as bound up with the fact that causes must be local in space and time to their effects. Thus, that all forces are constrained to operate by mediated contact concerns in the first instance the nature of the spatiotemporal arena within which things act. That the arena imposes limits on the kinds of inhabitants it can accommodate is what makes the constraint that all fundamental forces act by mediated contact qualify as an EFL (according to Hertz) and so as explanatorily prior to the constraint that all fundamental forces are inverse-square.

Of course, my purpose here is not to endorse Hertz's implicit conception of space as an inert stage having dimensions and other features that constrain the kinds of physical interactions there could be—just as I need not endorse the explanation that Hertz proposes (or even its explanandum). Rather, my purpose in this section is to understand the basis for the distinction between EFLs and EDLs. I think we can grant that the conception of space I have ascribed to Hertz is the kind of fact that could serve as such a basis in this case. But it could not play this role in every case—even in every case concerning spacetime geometry. For instance, it cannot ground the explanatory priority of the interval's invariance over the Lorentz transformations.

I therefore suggest that what makes one constraint an EFL rather than an EDL may have little to do with what makes another constraint an EFL rather than an EDL. This is not to say that the EFL/EDL distinction is groundless. Indeed, I have just given two examples of facts that might help to organize a given rung into EFLs and EDLs.

6. Conclusion

Explanations by constraint have been relatively neglected in recent literature on scientific explanation, especially as that literature has emphasized causal explanation. Explanations by constraint do not work by virtue of describing causal relations. Rather, explanations by constraint work by supplying information about the explanandum's relation to necessities that transcend ordinary causal laws. I have tried to unpack this idea and to show how it helps us to understand several notable examples of proposed explanations by constraint.

Some non-causal scientific explanations are not explanations by constraint. For instance, "dimensional explanations" work by showing how the law of nature being explained arises merely from the dimensions of the quantities involved. "Really statistical

²⁰ Callender (2005: 128) offers another case where the dimensionality of space seems to be recognized as taking explanatory priority over a feature of space's inhabitants, namely, that some forces are such as to permit stable orbits: "There is a strong feeling—which I think Russell, van Fraassen and Abramenko were all expressing—that stability is just the wrong kind of feature to use to explain why space is three dimensional.... The feeling is that stability... is simply not a deep enough feature to explain dimensionality; if anything these facts are symptoms of the dimensionality."

explanations” include explanations that explain phenomena by characterizing them as regression toward the mean or as some other canonical manifestation of chance. Elsewhere (Lange 2016) I describe these and other varieties of non-causal explanation. I argue that although these varieties work differently from one another and from causal explanations, they all are alike in certain respects (such as in their relations to certain properties being natural and to certain facts being coincidental). Therefore, despite their diversity, they all deserve to be grouped together—as explanations.

References

- Aharoni, J. (1965), *The Special Theory of Relativity*, 2nd edn. (Oxford: Clarendon Press).
- Baker, A. (2009), ‘Mathematical Explanation in Science’, *British Journal for the Philosophy of Science* 60: 611–33.
- Bartlett, D. and Su, Y. (1994), ‘What Potentials Permit a Uniqueness Theorem’, *American Journal of Physics* 62: 683–6.
- Barton, G. (1999), *Introduction to the Relativity Principle* (New York: Wiley).
- Berzi, V. and Gorini, V. (1969), ‘Reciprocity Principle and Lorentz Transformations’, *Journal of Mathematical Physics* 10: 1518–24.
- Bondi, H. (1970), ‘General Relativity as an Open Theory’, in W. Yourgrau and A. Breck (eds.), *Physics, Logic, and History* (New York: Plenum Press), 265–71.
- Bondi, H. (1980), *Relativity and Common Sense* (New York: Dover).
- Brading, K. and Castellani, E. (2003), *Symmetries in Physics: Philosophical Reflections* (Cambridge: Cambridge University Press).
- Braine, D. (1972), ‘Varieties of Necessity’, *Supplementary Proceedings of the Aristotelian Society* 46: 139–70.
- Brown, H. (2005), *Physical Relativity* (Oxford: Clarendon Press).
- Callender, C. (2005), ‘Answers in Search of a Question: “Proofs” of the Tri-Dimensionality of Space’, *Studies in History and Philosophy of Modern Physics* 36: 113–36.
- Earman, J. (1989), *World Enough and Space-Time* (Cambridge, MA: MIT Press).
- Eddington, A. (1920), *Space, Time and Gravitation* (Cambridge: Cambridge University Press).
- Hertz, H. (1999), *Die Constitution der Materie* (Berlin: Springer-Verlag).
- Lange, M. (2008), ‘Why Contingent Facts Cannot Necessities Make’, *Analysis* 68: 120–8.
- Lange, M. (2009), *Laws and Lawmakers* (Oxford: Oxford University Press).
- Lange, M. (2013), ‘What Makes a Scientific Explanation Distinctively Mathematical?’, *British Journal for the Philosophy of Science* 64: 485–511.
- Lange, M. (2014), ‘Aspects of Mathematical Explanation’, *Philosophical Review* 123: 485–531.
- Lange, M. (2016), *Because Without Cause: Non-Causal Explanation in Science and Mathematics* (Oxford: Oxford University Press).
- Lee, A. and Kalotas, T. (1975), ‘Lorentz Transformations from the First Postulate’, *American Journal of Physics* 43: 434–7.
- Lévy-Leblond, J.-M. (1976), ‘One More Derivation of the Lorentz Transformations’, *American Journal of Physics* 44: 271–7.
- Lewis, D. (2007), ‘Causation as Influence’, in M. Lange (ed.), *Philosophy of Science: An Anthology* (Malden, MA: Blackwell), 466–87.

- Mancosu, P. (2008), 'Mathematical Explanation: Why It Matters', in P. Mancosu (ed.), *The Philosophy of Mathematical Practice* (Oxford: Oxford University Press), 134–50.
- Mermin, N. D. (2009), *It's About Time* (Princeton: Princeton University Press).
- North, J. (2009), 'The "Structure" of Physics: A Case Study', *Journal of Philosophy* 106: 57–88.
- Owens, D. (1992), *Causes and Coincidences* (Cambridge: Cambridge University Press).
- Pincock, C. (2007), 'A Role for Mathematics in the Physical Sciences', *Noûs* 41: 253–75.
- Pincock, C. (2015), 'Abstract Explanations in Science', *British Journal for the Philosophy of Science* 66: 857–82.
- Salmon, W. (1998), *Causality and Explanation* (Oxford: Oxford University Press).
- Stachel, J. (1995), 'History of Relativity', in L. Brown, A. Pais, and B. Pippard (eds.), *Twentieth Century Physics*, volume 1 (College Park, MD: American Institute of Physics Press), 249–356.
- Steiner, M. (1978a), 'Mathematical Explanation', *Philosophical Studies* 34: 135–51.
- Steiner, M. (1978b), 'Mathematics, Explanation, and Scientific Knowledge', *Noûs* 12: 17–28.
- Watkins, J. (1984), *Science and Scepticism* (Princeton: Princeton University Press).
- Wigner, E. (1972), 'Events, Laws of Nature, and Invariance Principles', in *Nobel Lectures: Physics 1963–1970* (Amsterdam: Elsevier), 6–19.
- Wigner, E. (1985), 'Events, Laws of Nature, and Invariance Principles', in A. Zuchichi (ed.), *How Far Are We from the Gauge Forces* (New York: Plenum), 699–708.
- Woodward, J. (2003), *Making Things Happen* (Oxford: Oxford University Press).

2

Accommodating Explanatory Pluralism

Christopher Pincock

1. Introduction: Strong and Weak Explanatory Pluralism

A pluralist about X maintains that X s come in a variety of different types $X_1, \dots, X_n$. A pluralist about value, for example, points to different types of values and argues for their philosophical significance. My focus in this chapter is explanatory pluralism. To be an explanatory pluralist is to insist that explanations come in several types. Many discussions of explanatory pluralism consider only what could be called minimal pluralism. A minimal pluralist insists that a genuine explanation with certain virtues cannot be replaced by an explanation of another type with those very same virtues. This is how Lipton argues for a view that he calls “pluralism”. Scientific explanations of various sorts are needed, and so there is no point, e.g., in trying to reduce all causal explanations to micro-causal explanations: “A good scientific explanation sometimes requires macro causes, sometimes micro causes, and sometimes a combination of the two. When it comes to scientific explanation, we should be pluralists” (2008: 124). Lipton’s argument for pluralism supposes that these different types of explanations tend to realize different explanatory virtues such as strict necessity as opposed to “generality and unification” (2008: 122). What makes a scientific explanation of one sort good is often just not something that can be matched by an explanation of another sort. I take this minimal form of pluralism to be very plausible, and also easy to accommodate on a wide range of views about explanation.

A more controversial form of pluralism claims that for each genuine explanation E_1 of one type there simply is no genuine explanation E_2 of another type that incorporates, subsumes, or absorbs E_1 .¹ For one explanation to absorb another is for that explanation

¹ Cf. Reutlinger (2016). He argues that the pluralist must show that there is no theory that covers all explanations. I believe that this places an unfair burden on the pluralist as they must argue that explanations of different types resist any unified theoretical treatment.

to have the other one as a part. Schematically, if E_1 takes the form of C standing in relation R to E , then it will be absorbed by E_2 when E_2 takes the form of C standing in relation R to E along with other facts, such as that D stands in relation R to C . Exactly what this comes to depends on whether one adopts an ontic or an epistemic approach to explanation. An ontic approach identifies both the object of the explanation and the explanation itself with facts. What makes some facts explain another fact is a feature of the world as it is independent of human agents. By contrast, an epistemic approach adds an essential reference to human agents and their knowledge states. So in order to say what makes some facts explain another fact, an epistemic view will add additional tests tied to the states of the agents doing the explaining.

Explanatory pluralism requires that explanations come in different types. On an ontic interpretation, what this means is that there is an explanation E_1 of type T_1 of object of explanation O , and the facts making up E_1 are not a part of any more encompassing explanation of any other type.² An epistemic approach will say something quite similar except this approach can use knowledge states as well to block one explanation from being absorbed into another. As I will discuss in section 2, one type of explanation is causal explanation. So the explanatory pluralist is committed to there being explanations that are not part of any causal explanation. But each type of explanation may have interesting internal relations. For example, one causal explanation may be subsumed under another causal explanation.

On both an ontic and epistemic view, a genuine explanation will require facts that bear the right relation to the fact being explained, and each of these facts will typically be represented by a true proposition. Two sorts of non-minimal explanatory pluralism are examined in this chapter. Strong explanatory pluralism maintains that some explanatory targets have genuine explanations of different types. That is, for some object of explanation O , both E explains O and F explains O and these explanations are of different types. There are two ways to show that alleged explanations of different types are actually of the same type.³ Either argue that one explanation actually includes the other or that both are included in a third more encompassing explanation. Consider, for example, two causal explanations of an event. If some light bulb turned on because an electrical current was running through a circuit, then that constitutes one causal explanation for why a light bulb went on. But another explanation of the same type is that a switch was flipped, and allowed the current to run through the circuit, and this turned the light on. This second explanation subsumes the first explanation, and this shows that they are of the same type. There are also cases of genuine explanations of the same target where neither includes the other, but both are subsumed by some third explanation. That one switch was flipped explains why at least one light bulb went on and that another switch was flipped also explains why at least one light bulb went on.

² I suppose here that O is some fact.

³ These are two sufficient conditions for being of the same type. Necessary and sufficient conditions for being of the same type are given in section 2.

But that the department head ordered that more lights be turned on explains both why the first switch was flipped and why the second switch was flipped, and so why at least one light bulb went on. This shows that these two explanations of that target are of the same type.⁴

For strong explanatory pluralism to be true there must be distinct types of explanation. In section 2 I introduce three types of scientific explanation: causal, constitutive, and abstract. A causal explanation cites the causes of the phenomenon being explained, while a constitutive explanation indicates what composes the phenomenon and how that composition makes the phenomenon obtain. In addition, I argue that there is a third type of explanation that I call abstract. An abstract explanation points to certain abstract characteristics of the system that make the system have certain features. If these are all genuine explanations, and they apply to the very same target, then strong explanatory pluralism is vindicated. There will be explanations of some target phenomenon that are free-standing of one another in the sense that there is no potential to absorb any two of them into some more encompassing explanation. An explanation of a given type, when it is found, provides something that no explanation of any other type can offer.

Strong explanatory pluralism can be contrasted with a weaker explanatory pluralism that merely insists that explanations come in different types. Weak explanatory pluralism does not require that there is some single target that is explained by explanations of different types. It is consistent with this possibility, but also consistent with each type of explanation having its own special sort of explanatory target. For example, one might think that there is a special sort of explanation found in pure mathematics. The object of these explanations is the truth of some mathematical theorem. A purely mathematical explanation of the truth of some theorem might involve a proof that has special characteristics that distinguish it from other proofs that merely show that the theorem is true. One could believe in this type of explanation and yet remain a weak explanatory pluralist. This position would insist that there are no purely mathematical explanations of non-mathematical targets. There is thus no overlap between the objects of these mathematical explanations and the other types of explanation, such as causal explanations. A strong explanatory pluralist denies that the objects of explanations are sorted into these disjoint families. Again, there are some targets of genuine explanations that have two or more types of explanation.

Both the weak and the strong explanatory pluralist face a general challenge that arises for any form of pluralism. Suppose we have a list of different types of explanations such as causal, constitutive, and abstract. The pluralist then faces an unappealing dilemma. Either the members of this list have nothing in common or they have

⁴ Brigandt (2013) deploys a similar contrast between strong and weak explanatory pluralism. His argument for strong explanatory pluralism concerns explanatory models that “make jointly incompatible idealizations (necessitated by different explanatory aims)” tied to different research programs (2013: 88). This is not the argument I develop here, but I must reserve engaging with this argument for future work. See also Woody (2015) and Potochnik (2015) for related arguments.

something in common. If the members of this list have nothing in common, then it is hard to say why they are actually types of explanation. They may be something more generic such as facts, but they lack any common core that unites them all as explanations. However, if the members of the list do have something in common, and if this is to illuminate how they are all types of explanation, then it is not clear what kind of pluralism can be maintained. A weak pluralist points to mathematical explanations of mathematical theorems and causal explanations of physical events, and supposes that they are all explanations despite their different targets. The strong pluralist adds that some causal explanations are explanations of the very same things as some constitutive or abstract explanations. Either way, it remains unclear how all these accounts can be explanations and yet fall into irreducibly different types. The pluralist owes us a discussion of what all explanations have in common and what nevertheless divides these explanations with this common feature into distinct types. Otherwise the common feature threatens to unify explanations into a single type and pluralism of any form is blocked.

In the rest of this chapter I argue for three claims. First, the diversity of explanations found in scientific practice mandates some form of explanatory pluralism. Second, the most promising form of explanatory pluralism is a version of weak explanatory pluralism that insists that the target of each explanation is a contrast of the form *P* rather than *Q*. Third, this flavor of explanatory pluralism fits with a version of an ontic approach and a version of an epistemic approach, but both views face challenges. The ontic approach has difficulty making sense of contrastive facts. The epistemic view can make sense of the explanation of contrasts by appeal to the knowledge states of agents. But it remains unclear how either approach can vindicate the value that scientists place on finding explanations as opposed to merely true descriptions of phenomena.

2. Three Types of Explanation

Cases drawn from scientific practice can be used to motivate explanatory pluralism. Here I will sketch three cases that support the conclusion that explanations come in at least three types. These are causal explanation, constitutive explanation, and abstract explanation. A case that illustrates these three types is the board of directors of some organization that is made up of people, all of whom are bald.⁵ Suppose that we aim to explain why all the directors are bald. There are three kinds of explanation of this general fact. As we will see, constitutive explanations and abstract explanations are different from causal explanations in virtue of containing a special sort of non-causal relation. A constitutive explanation makes essential use of part/whole relations. An abstract explanation uses the fact that one thing instantiates another. For example, an

⁵ This case is emphasized for different purposes in Hempel (1965). As will become clear, my treatment of this example is influenced by the classic discussion of Garfinkel (1981) and the more recent Haslanger (2016).

abstract geometrical structure may be instantiated by a physical system, but the structure is not a part of the system. By contrast, a causal explanation exploits only causal relations. Using these assumptions, I will argue that causal, constitutive, and abstract explanations are of different types. The distinctive non-causal relations found in constitutive and abstract explanations block any attempt to subsume them under causal explanations. For similar reasons, we can neither subsume a constitutive explanation under an abstract explanation nor subsume an abstract explanation under a constitutive explanation. A necessary and sufficient condition for being of the same explanatory type, then, is that two explanations exploit the same explanatory relations. If explanation A uses relation R and explanation B uses relation S, then A and B are of different types.

This way of dividing up explanations into types is further motivated by the widely accepted point that adding more facts can spoil an explanation. Suppose, for example, that A stands in relation R to B, and that this fact is a causal explanation of B. It does not follow that the combined fact that C stands in relation S to B and that A stands in relation R to B is also an explanation of B. This “non-monotonic” aspect of explanation holds even when it is the case that the fact that C stands in relation S to B alone is an explanation of B. Combining explanations need not preserve there being an explanation.

One genuine explanation of the fact that the board of directors are all bald is the votes of the membership that elected each director. In a series of elections, first A got the most votes, then B got the most votes, and so on until all the elections are covered. If we add that A is bald, B is bald, and so on until each director is mentioned, we have an explanation of why all of these directors are bald. On many views of causal explanation, this amounts to a genuine causal explanation. Here I suppose that Woodward has developed an adequate account of causal explanation, and our sketch certainly counts as a causal explanation by Woodward’s lights (Woodward 2003). Woodward emphasizes the need to say how the actual situation would have differed if at least one parameter is varied, while others are held fixed at their actual values. Woodward adds the restriction that a parameter is varied by an “intervention”. This limits his test to cases where a causal relation obtains. In the board of directors case, a change in the votes during the election that actually elected A would have resulted in the election of a rival candidate Z. If we suppose that Z is not bald, then this change in the votes would have made it the case that some of the board members are not bald. For Woodward, this amounts to a causal explanation of why all the board members are bald.

A second explanation notes that A is bald because he lacks sufficiently many hairs on his head. This second explanation would point to a similar condition for B and the other directors. The distinctive feature of this explanation is that it cites the composition of A, B, and the rest in the sense that the lack of hairs are parts of these people. If A was composed differently, and as a result had hairs on his head, then he would not be bald. And if he were not bald, then it would not be the case that all the members of the board were bald. When an explanation appeals to the parts of the phenomenon being explained, then I will call it a constitutive explanation.

My argument that this explanation is of a different type than any causal explanation is that this explanation deploys the part/whole relation in an ineliminable way. So, no causal explanation can fully absorb this constitutive explanation. However, it is not immediately clear that this argument works. It might seem that Woodward's notion of an intervention is flexible enough to accommodate whatever is genuinely explanatory in this explanation. If so, then the explanatory role of the part/whole relation is minimal.

Recent discussions of Woodward's notion of an intervention have highlighted this issue in connection with cases where there are non-causal dependencies between variables (Shapiro and Sober 2007; Woodward 2015). In Woodward's example, a person's level of cholesterol TC is the sum of their LD and HD levels of cholesterol. Woodward claims that TC and LD stand in a non-causal relation of "definitional dependence" (2015: 327). He uses this relation to understand other non-causal relations of dependence, especially supervenience relations. In the cholesterol case there is no "relevant" intervention on LD that fixes TC at its actual value. In our case, we can suppose that a person's baldness B is a variable with values 1 for "bald" and 0 for "not bald", and also that the density D of the hairs on their head determines their baldness. If D is greater than some threshold, then B = 1. If D is below that threshold, then B = 0.⁶ But the value of D constitutes the person's baldness, rather than causing it. An explanation that proceeds through this sort of link is thus quite different than an ordinary causal explanation. The part/whole relation is not eliminated or replaced by wholly causal relations. In this sense, then, my original argument stands.

A third type of explanation of the baldness of the board of directors is available. This is the structural, or what I will call "abstract", explanation. Suppose that the elections occur in a highly sexist society that gives men many more opportunities for professional advancement. This sexism structures the election of the board members in such a way that it nearly guarantees that all the board members are men of a certain age. If we suppose also that baldness is much more common among men of that age than among women or younger men, then we have a distinct structural explanation of the makeup of the board of directors. There are abstract features of the whole organization and the society that it is a part of that are highly conducive to this outcome.⁷

The special feature of this abstract explanation is that it abstracts away from the constitutive features of the board members. There is nothing special about A, according to this explanation, that made him get elected to the board. For if A had been sidelined through some personal misfortune, and not had the opportunity to run for the board, then the abstract structure of the whole system is such that another candidate A' would have run in his place. And given the character of this system A' is overwhelmingly likely to have been an older man. This shows the gap between our constitutive explanation and our structural explanation. No appeals are made to the particular elements of the system or their internal constitution.

⁶ Here I ignore the complications associated with the vagueness of this predicate.

⁷ This is not the same as Jackson and Petit's notion of program explanation. See Pincock (2015: 871–4) for a discussion of the differences.

It remains debatable, of course, how different the structural explanation is from any causal explanation.⁸ One might, for example, propose a Woodward-style interpretation of the structure of the society in question. It is certainly true that if this sexist structure were changed to a more egalitarian structure, then the composition of the board with respect to its baldness is highly likely to be changed. What is less clear is whether this sort of change should be thought of as a change in the value of a variable. In the constitutive case we saw that it was important to recognize two sorts of links between variables, namely causal and constitutive. In the structural case, the structural features serve as a kind of background against which causal links between variables are established. The structure is thus taken for granted in a way that distinguishes it from any particular causal variable. This suggests that we should contrast what the structure enables or inhibits with what happens within that structure. A change in structure can have dramatic effects that will complicate any notion of the intervention on a structure.

The point can be further supported by appeal to the instantiation relation. An abstract structure is instantiated by a more concrete system. In an abstract explanation the instantiation relation plays a part in the explanation. In our baldness case, there is a complex social structure that maps out the abstract network of gender relations in our society. This structure is instantiated in our society. This fact forms a central part of the abstract explanation for why all the board members are bald. The instantiation relation here cannot be replaced by causal or constitutive relations as this abstract structure neither causes nor constitutes the network of gender relations found in our society. I conclude that this abstract explanation cannot be absorbed into an explanation of either of the other two types.

Our discussion of the board of directors case shows the need for at least the types of explanations that I have called causal, constitutive, and abstract. Two more mathematical cases can be used to make the same point (Pincock 2007, 2015). The residents of Königsberg wondered why they had failed to make a circuit of their city that involved crossing each of its seven bridges exactly once. A causal explanation of this pattern could appeal to each of the attempted circuits and indicate how it had failed by either crossing a bridge more than once or missing a bridge. Each failure has its cause, and this cause can be given a Woodward-style analysis. For example, Wilhelm's failure to complete a circuit arises when he crosses the western most bridge twice. He would not have failed this way if he had turned left rather than right at one point in his journey.⁹ A constitutive explanation could appeal to the material that made up the bridges and the people making the crossings. Wilhelm would not have failed in the way he did if he

⁸ Although Haslanger draws attention to the importance of structural explanations and interprets them in terms of the instantiation of abstract structures, she also appears to view them as a special kind of causal explanation. In particular, Haslanger relates her structural explanations to Dretske's "structuring causes" (2016: 120).

⁹ One might worry that this causal explanation does not explain the very same fact as the constitutive and abstract explanations. I develop this point in section 4 using contrastive facts.

had been paralyzed. Finally, an abstract explanation could appeal to the structure of the bridges. This structure ensures that no attempted circuit would be successful.

An even more mathematical example concerns the laws for how soap-film surfaces meet in stable soap-film configurations. Plateau noticed certain patterns to these meetings that he codified into three laws. A causal explanation of this pattern would indicate the mechanism through which these systems minimize their surface area, subject to the constraints imposed. A constitutive explanation could summarize the spatial arrangement of the parts of each such system and show how they conform to Plateau's laws. Finally, an abstract explanation would show how the patterns found by Plateau follow from a more general mathematical structure. Any instance of that mathematical structure would conform to Plateau's laws.¹⁰

3. Ontic Accounts

Causal, constitutive, and abstract explanations are different types of explanations.¹¹ It looks like the same fact is being explained across types and so our cases appear to support what I have called strong explanatory pluralism. Ontic accounts that identify explanations with facts have great difficulty in accommodating strong explanatory pluralism. In the remainder of this section I will consider two ontic attempts to accommodate this kind of pluralism. The first attempt generalizes Woodward's notion of an intervention to cover all three types. The second attempt deploys the concept of ontological dependence to make sense of each of these explanations. Both attempts face the same problem. They wind up with such a weak common feature among explanations that they lose a substantial account of what makes explanations valuable. For this reason, these proposals cannot distinguish explanations from non-explanations.

We have already seen that Woodward's notion of a causal relation tied to interventions is too narrow to include constitutive part/whole relations. The same point holds for structural instantiation relations, as Woodward notes in passing (2003: 220). However, one could try to identify a more generic notion of "difference making" that includes all three of these explanatory relations. Woodward himself talks of "what if things had been different". It might seem that a broader modal test could identify what our three explanatory relations had in common. But this common feature would not undermine explanatory pluralism as the more specific characteristics of these relations could still play a role in individuating types of explanations.¹²

¹⁰ See Pincock (2015), Saatsi (2016), and Baron et al. (forthcoming) for more discussion of mathematical explanations of physical phenomena. Andersen (forthcoming) develops a very different picture of these cases. She uses a notion of a model "holding of" a system to motivate strong explanatory pluralism. I unfortunately lack the space to discuss this important argument here.

¹¹ These types of explanation have some affinity to Aristotle's efficient, material, and formal causes, respectively. I defer to future work an investigation of a modern analogue of Aristotelian final causes in the explanation of human action.

¹² See especially Saatsi and Pexton (2013), Rice (2015), and Reutlinger (2016).

For the board of directors case, the causal explanation meets Woodward's more demanding intervention test: there is an intervention on the variable that reflects the vote that elected A such that Z is elected instead. This change results in a change in the baldness state of the board, as we supposed that Z is not bald. By passing this more demanding test, the causal explanation also passes a more generic modal test: it tells us how things would have been different, namely how the baldness state would have changed if the vote had gone that way. So far, so good. A similar pattern obtains for the constitutive explanation. Now we explain the baldness of the board via the composition of its members and their internal constitution. The part/whole relation here does not pass Woodward's intervention test, but it does pass the more generic modal test: if A had been constituted differently, so that A was not an older male, but was instead a woman, then A would not have been bald. So the baldness state of the board would have changed if A's internal constitution had been changed. Finally, consider the structural explanation that appeals to the instantiation of a sexist social structure. If the system had not instantiated this structure, but instead instantiated the structure of an egalitarian society, then the board would no longer have its baldness state. The structural explanation also indicates what would have been different, but now via its instantiation relation.

The current proposal, then, is that each type of explanation explains by deploying a relation that indicates how things would have been different if various changes had been introduced into the actual board of directors system. What varies across types is *how* this relation gives this modal information. That is why there is a genuine form of pluralism. But there is still a unified core to this class of genuine explanations: if modal information is provided, then one has a genuine explanation.

One problem with this proposal is that it is too flexible.¹³ There are simply too many cases where an account that fails to be a genuine explanation deploys a relation that provides the right kind of modal information. Many of these cases can be found in classic objections to Hempel's D-N account of explanation. Consider, for example, the attempt to explain E using C where there is no causal link from C to E, and yet C and E are highly correlated due to some common cause F. Thunderstorms are caused, in part, by a drop in atmospheric pressure. And a drop in atmospheric pressure also causes a barometer to show a lower reading. This generates a strong correlation between a barometer showing a lower reading and a thunderstorm occurring. If a scientist proposed that the barometer's lower reading explained the thunderstorm, then this proposed explanation would be rejected as not genuine. However, this proposed explanation certainly does convey the right kind of modal information. It says how things would have been different: if the barometer had not given the lower reading, then the thunderstorm would not have occurred. This shows that merely conveying modal information is not sufficient for providing a genuine explanation.

¹³ Another worry is that it fails for cases that involve pure mathematics. See Baron et al. (forthcoming) for a recent discussion. I am grateful to an anonymous referee for emphasizing this problem.

Another proposal along these lines is to require that the modal information be conveyed by appeal to one of the following relations: (i) causal, (ii) constitutive, or (iii) structural instantiation.¹⁴ The proposed barometer explanation fails this more demanding test because that account did not link the barometer to the thunderstorm by a causal, constitutive, or structural relation. This revised modal proposal faces two problems. First, it does not clarify why it is *these* three relations that are needed for an explanation. If a new relation was considered as a supplement to this list, then how are we to tell that it could or could not generate genuine explanations? If providing modal information is not sufficient, it is unclear why providing modal information by one or the other of these relations is sufficient. Second, there are counterexamples like the barometer case that provide modal information via one of these relations, but yet are not genuine explanations. Consider, for example, a failed constitutive explanation of the board of directors' baldness. It may be the case that any alteration of a board member's genetic makeup that is sufficient to lower their risk of heart attack would also lower their baldness. So, we can truly say that were some board member to have a lower risk of heart attack, then the board would not be composed entirely of bald people. This proposed explanation conveys modal information by appeal to a constitutive relation that obtains in the actual board, and yet it is not a genuine explanation. If this strategy is to accommodate explanatory pluralism, then a tighter set of conditions must be imposed.

A modal strategy tries to accommodate explanatory pluralism by tying each genuine explanation to a modal fact. A distinct ontic strategy is to focus instead on relations of ontological dependence. As emphasized by Fine, Koslicki, and others, ontological dependence relations may obtain even in the absence of the usual modal facts. The set whose only member is the number 3, for example, may be said to ontologically depend on the number 3 despite the necessary existence of both the set and the number 3. So it might seem promising to ground a form of explanatory pluralism on the obtaining of an ontological dependence relation. This is Koslicki's suggestion in her paper "Varieties of Ontological Dependence":

[...] an explanation, when successful, captures or represents [...] an underlying real-world relation of dependence of some sort which obtains among the phenomenon cited in the explanation in question [...] If this connection between explanation and dependence generalizes, then we would expect relations of ontological dependence to give rise to explanations within the realm of ontology, in the sense that a successful ontological explanation captures or gives expression to an underlying real-world relation of ontological dependence of some sort.

(Koslicki 2012: 212–13)

There is thus a list of dependence relations that includes (i) causal, (ii) constitutive, and (iii) structural instantiation. A genuine explanation of E in terms of C involves linking C to E by one of these dependence relations. This dependence need not involve any modal information, and so the presence or absence of modal features is not decisive in the evaluation of the proposed explanation. Instead, what is decisive is whether or not

¹⁴ A modal approach could of course be developed in other ways. Reutlinger (2016) clearly recognizes the worry raised in the last paragraph.

this special sort of relation obtains. Our causal explanation explains by citing the causal relation between the vote and A's presence on the board. The constitutive explanation explains via the constitutive relation that obtains between A's hairs and A's baldness. Finally, the structural explanation functions by appeal to the instantiation relation that obtains between the abstract sexist structure and the society which instantiates it.

One worry about the dependence proposal is that it is hard to figure out what all these ontological dependence relations have in common. One suggestion is:

(*) that E ontologically depends on C just is that C makes E obtain.

This natural suggestion faces an overdetermination problem if we add the suppositions that there are distinct types of dependence relation and only one way for something to be made to obtain. Consider, again, the fact that all the members of the board of directors are bald. On the dependence proposal, this fact is explained in three different ways tied up with causal, constitutive, and structural dependence. Using (*), if the baldness state depends on its causes, then these causes together make the baldness state obtain. But equally, via (*), if the baldness state depends on its composition, then its composition makes the baldness state obtain. A similar point holds for the structural instantiation relation. The problem now is that there are three different types of facts, each of which serves to make the baldness fact obtain. How can this be? The dependence proposal must be revised to allow that each dependence relation makes a fact obtain in its own way. There is no competition between these ways and so no risk of overdetermination.

At this point the dependence proposal takes on a somewhat mysterious aura. Explanations explain because they involve these relations and these relations are significant because they make facts obtain, but each type of relation works differently and so can make a fact obtain in a different way. Again we face the problem of saying why certain relations make the list of dependence relations while others are excluded. It may just be a metaphysically primitive feature of the world. But if it is just a primitive feature of the world, then this strategy for accommodating explanatory pluralism leaves us with little recourse for resolving debates about explanation. Someone may propose, for example, that in addition to the way that wholes constitutively depend on their parts, there is also a way that parts holistically depend on the wholes they are a part of. This means that there are "holistic" explanations over and above the causal, decompositional, and structural explanations already considered. How can an advocate of our revised dependence proposal combat this suggestion or any other suggestion? Partly for this reason, we lose any link to the value that scientists place on having explanations. If we do not understand what makes a relation a dependence relation, then we also lack an understanding of what makes something an explanation. But scientists do value explanations, and so we must hope that there is some feature that all explanations have in common that makes the quest for explanation coherent. So far we have not found any way to do this consistent with strong explanatory pluralism.¹⁵

¹⁵ An ontic view of explanation could add on a further account of the cognitive state known as understanding. This appears to be Strevens's strategy for making sense of explanatory pluralism.

4. Contrastive Facts and a Viable Ontic Strategy

An ontic approach to explanation identifies explanations with facts. These facts explain some other fact according to the sort of relation emphasized by that specific flavor of ontic approach. So far we have considered two unsuccessful ontic attempts to making sense of strong explanatory pluralism. A modal approach zeros in on how some facts are responsible for the modal features of some actual fact. A dependence theory instead posits basic dependence relations that connect the facts doing the explaining and the fact being explained. The overly flexible character of the modal approach showed the need for a tighter connection between explanations and their targets. But the dependence approach faced an overdetermination problem and resolved it by supposing various ways that some facts could make a fact obtain.

A shift to weak explanatory pluralism makes room for a resolution of some of these problems. The overdetermination problem arises only because E explains O and F explains the very same O. If we could somehow distinguish the fact that E explains from the fact that F explains, then there would no longer be any obstacle to maintaining that E makes O obtain and also that F makes O' obtain. One promising way to do this for the cases we have considered is to suppose that the object of these explanations is a contrastive fact. In our discussion so far we have operated with the basic contrast between all the members of the board being bald and it not being the case that all the members of the board are bald. But we could add that there is actually a richer space of contrastive facts in play here, and that an explanation of a given type is suited to explain only one kind of contrastive fact. In this way, the objects of these explanations would themselves be sorted into disjoint types that reflect the types of the explanations. On this position, only weak explanatory pluralism obtains as there is no object that is explained by explanations of different types.¹⁶

An ontic approach can sidestep the overdetermination problem, then, by recognizing only weak explanatory pluralism, and one way to do that is to finely individuate the objects of explanation as contrastive facts. The same maneuver can be used to try to address the other worry about ontic approaches to pluralism. This is that there is no account of why certain relations give rise to explanations and others do not. With contrastive facts at its disposal, the ontic approach can use the character of the contrasts to motivate the explanatory relations and distinguish them from the non-explanatory relations.

To see how this might work, let us reconsider the board of directors case, but now in terms of various contrastive facts. We supposed that A got more votes than another candidate Z, and that A was bald and Z was not bald. The contrast then is between the board of directors (including A) all being bald rather than the board of directors (including Z) nearly all being bald. This contrast is well explained by the causal

¹⁶ Hitchcock (2012) argues for different types of explanation and that the object of each explanation is a contrast. However, he does not claim that each contrast is apt to be explained by at most one type of explanation. He seems to endorse the dependence proposal discussed in section 3. (See especially 2012: 26.)

explanation that cites the votes in the election that gave A more votes than Z. But, crucially, this very contrast is not explained by either the constitutive or the abstract explanation. The constitutive explanation considered the parts of each of the actual board of directors and indicated how the actual parts gave rise to the baldness of each. This has no bearing on how Z could have become a board member. Similarly, the abstract explanation cited the instantiation of a sexist social structure. This sexist social structure has no tie to the contrast between A and Z being on the board as that structure is being held in place across this contrast.

What related contrasts, then, are apt to be explained by a constitutive or an abstract explanation? Consider the contrast between the board of directors all being bald rather than some of those very board members not being bald. To explain this we cannot cite the votes that elected the actual board members. We must instead consider the internal constitution of some of those board members. Clearly, if A's internal constitution had been different, such that he had more hairs on his head, then he would not be bald. So we see that a constitutive explanation is well-suited to explain this contrast. The contrast, in effect, holds fixed the chain of events leading up to these people being on the board, but requires us to consider changes in the people's internal constitution. This is why a constitutive explanation is appropriate and no causal explanation can succeed.

The abstract explanation is designed to explain the following contrast: the board of directors all being bald rather than being reflective of the rate of baldness of the general population. Let us suppose that 25 percent of the population is bald. This contrast can be explained by giving some basis for the gap between the 100 percent baldness of the board and the 25 percent baldness of the population that the board is drawn from. The fact that the society instantiates a sexist social structure does explain this contrast as it classifies the actual society in a way that shows how the two percentages could diverge so sharply. There is a kind of top-down structuring to the events leading up to these board members all being bald. By contrast, in other societies where a different, more egalitarian social structure is instantiated, more of a match between the population and the board is to be found. Neither the causal explanation nor the constitutive explanation fits this contrast. The causal explanation considers how causes operate within the given social structure and so does not factor in what is due to that structure itself. The constitutive explanation varies only the internal constitution of the actual board members, and so also does not consider the role of the abstract social structure.

Schematically, then, we have three kinds of contrastive facts and we can suppose that there is something about the kind of contrastive fact that makes it well-suited to be explained only by an explanation of a single type. Roughly, when a contrast is tied to a difference that could have been made through causes changing events, while fixing the constitutive character and the broader abstract structure, then a causal explanation is mandated. When a contrast relates to a change in the internal constitution of one or more elements, while not varying the causes between events or the broader abstract structure, then a constitutive explanation is required. Finally, when a contrast invokes a difference between types of systems, then only an abstract explanation will cite the

right kind of factor that is responsible for those differences across systems. Looking to the operations of causes or the internal constitution of the elements of the actual system will fail to make sense of that sort of contrast.¹⁷

An ontic account that embraces this kind of weak explanatory pluralism thus avoids the overdetermination problem and is able to motivate their list of explanatory dependence relations. The relations that figure in explanations naturally fall out of the character of the contrasts being explained. Does this show that there is an ontic route to accommodating explanatory pluralism?

5. A Viable Epistemic Strategy

Lipton has done the most to distinguish contrastive facts from other kinds of facts, and to make contrastive facts the proper objects of many scientific explanations (Lipton 2004, 2008). Lipton notes that we sometimes explain contrastive facts of the form *P* rather than *Q* and uses this practice to argue that a contrastive fact is distinct from any other kind of fact. Consider, for example, the attempt to reduce each contrastive fact *P* rather than *Q* to a conjunction of *P* and not-*Q*. Lipton rejects this reduction by pointing out that some explanations of the contrastive fact fail to be explanations of the conjunction. Here he uses the assumption that an explanation of a conjunction must explain both of the conjuncts. But in a classic case like Jones rather than Smith having paresis, one genuine explanation is that Jones rather than Smith has untreated syphilis. This explanation of the contrast fails to explain the conjunction that Jones has paresis and Smith does not because untreated syphilis rarely leads to paresis. So the contrastive fact is different from the conjunctive fact by a principle of indiscernibility of identicals. One cannot reply to Lipton's argument by saying that this fact is after all a genuine explanation of the conjunction, but just a poor explanation. For even if it is a poor explanation of the conjunction, the fact is a good explanation of the contrast. This shows that the contrastive fact and the conjunctive fact are different.

Lipton also offered a helpful difference condition that gives a necessary condition on a causal explanation of a contrastive fact:

To explain why *P* rather than *Q*, we must cite a causal difference between *P* and not-*Q*, consisting of a cause of *P* and the absence of a corresponding event in the case of not-*Q*. (Lipton 2004: 42)

We have seen this principle at work in our causal explanation of the baldness of the board of directors. One cause of the board of directors being bald (with *A* a member) rather than not bald (with *Z* a member) is the vote that elected *A* rather than *Z*. That vote caused *A* to be elected, and it corresponds to the absence of *Z*'s getting more votes.

¹⁷ Sober (1986) and Hitchcock (2012) independently suggest that contrasts have presuppositions. The character of these presuppositions may explain why only one type of explanation works for a given contrast. However, Sober and Hitchcock focus on causal explanation and do not seem to have extended this insight to non-causal cases.

Lipton unfortunately does not generalize his difference condition to non-causal explanations of contrasts. This is somewhat surprising as he argues for the existence of non-causal explanations. We can fill this gap by recasting Lipton's condition in more generic terms. Now we must speak of "explanatorily relevant factors" instead of causes, and also not suppose that these factors pertain just to events. As we have seen, the explanatorily relevant factors may relate to an entity's internal constitution or to the abstract structure instantiated by the system. With this wider range of cases in mind, Lipton's principle becomes:

To explain why P rather than Q, we must cite an explanatorily relevant difference between P and not-Q, consisting of a feature of P and the absence of a corresponding feature in the case of not-Q.

We considered a constitutive explanation of the contrast between all the board of directors being bald rather than those very same members not all being bald. This explanation cited A's internal constitution and its tie to being bald. Similarly, we reviewed the abstract explanation of the contrast between all the board of directors being bald rather than reflecting the rate of baldness of the general population. This explanation turned on the sexist social structure that was instantiated in the actual society. Its presence directed older men, and so people who were highly likely to be bald, to the board, while the presence of a more egalitarian social structure leads to more representative outcomes for boards in other sorts of societies.

It remains unclear how to approach explanations of contrastive facts on an ontic view of explanation. Recall that an ontic view identifies explanations with facts. So even though there are true propositions for each explanation, these propositions are not integral to the explanation when properly conceived. The difficulty is that the contrastive facts and the explanations we have arrived at are often tied up with the interests of the scientists investigating the world or other contextual factors over and above the facts themselves. The contextual aspects of contrastive explanation are emphasized by van Fraassen and Garfinkel, and even Lipton seems willing to concede their importance:

What makes one piece of information about the causal history of an event explanatory and another not? The short answer is that the causes that explain depend on our interests. But this does not yield a very informative model of explanation unless we can go some way towards spelling out how explanatory interests determine explanatory causes. One natural way to show how interests help us to select among causes is to reveal additional structure in the phenomenon to be explained, structure that varies with interest and that points to particular causes. The idea here is that we can account for the specificity of explanatory answer by revealing the specificity in the explanatory question, where a difference in interest is an interest in explaining different things. (Lipton 2004: 33)

Lipton's discussion suggests a world of facts with an overwhelming array of explanatorily relevant factors for each of these facts. An agent investigating the world then selects out a contrast as her object of explanation. Once this selection is made, and only

once this selection is made, there is a determinate answer for whether any proposed explanation of this contrast is a genuine explanation of that contrast. This is partly because of Lipton's difference condition. But we could add that the contrast that has been selected is only able to be explained by one type of explanation. So the selection of the contrast not only cuts down the number of explanatorily relevant factors, but also specifies that only one type of factor is relevant.

The viability of an ontic account of explanation turns on its making sense of this selective function of agents.¹⁸ A non-ontic, more epistemic alternative could match many of the advantages of an ontic account by focusing on questions of knowledge. On the ontic view, the contrastive fact that is the object of explanation is a genuine fact that emerges somewhat mysteriously out of the non-contrastive facts that obtain in a given situation. Whenever P and not-Q obtain, then P rather than Q obtains, although these are different facts. Accommodating weak explanatory pluralism has led the ontic account to privilege these contrastive facts as the objects of many scientific explanations. An epistemic alternative takes a different view of these contrastive facts. On this alternative approach, it is agents who know the conjunctive fact that P and not-Q and this knowledge is then presupposed in any legitimate explanatory question. When an agent knows the conjunctive fact, then they are able to pose the explanatory question "Why P rather than Q?" However, on the epistemic approach there is no need to posit any further contrastive fact. Instead, it is the agent's knowledge and their interests together that generate a legitimate question. The legitimacy of the question is established by factors beyond the obtaining of the conjunctive fact.

The epistemic alternative is non-ontic because it invokes factors beyond the facts in the world by themselves when determining whether or not something is a genuine explanation. These factors pertain to knowledge states and other states of the agents investigating the world. It is partly in virtue of these factors that an explanatory question is legitimate. One worry that is often raised against this sort of proposal is that it makes the existence of genuine explanations too closely tied to features of agents. As a result, it looks like we must index the genuineness of an explanation to a time, person, or research community. Newton had a genuine explanation of the fall of bodies on Earth *for Newton*, while Einstein had a genuine explanation of the fall of bodies on Earth *for Einstein*.

Given what we have seen so far, the epistemic account sketched here is not vulnerable to this form of relativism. For, just as with the ontic account, we can suppose that a contrast is apt to be explained by only one type of explanation. And, with Lipton, we can suppose that which explanations of this type are genuine is fixed only by the contrast and the facts that obtain in the world. Contextual factors like states of knowledge and interests do serve to determine which explanatory questions are legitimate

¹⁸ My narrow concern is quite different from Wright's sweeping attack on ontic approaches. Essentially, Wright assumes that "explaining designates a processual activity, which static or inert objects like sundials are incapable of performing" (2015: 29). But a defender of an ontic approach can and should distinguish the act of explaining from the explanation itself. Similarly, one should distinguish the act of pointing from the object that is pointed out.

for which agents. But the role of the context is limited to just this step. Once the explanatory question is in place, only certain explanations count as genuine, and what makes them genuine is that they reflect actual presences and absences of the right sort of explanatory factors.

This epistemic approach can endorse Woodward's picture of the very limited role of "pragmatics" in a theory of scientific explanation: "what we want to explain—the particular explanandum we want to account for—often depends on our interests or on contextual or background factors" (Woodward 2003: 229). Unlike Woodward, though, this epistemic account makes explicit how some knowledge states figure into the selection of a legitimate object of explanation.

We have arrived, then, at two somewhat equally matched strategies for accommodating explanatory pluralism. Both the ontic view and the epistemic view first retreat to weak explanatory pluralism by finely individuating the objects of explanation in terms of contrasts of the form *P* rather than *Q*. The ontic view supposes that there is a contrastive fact in the world, and that its internal character makes it apt to be explained by only certain kinds of other facts in the world. The epistemic approach instead adds an account of legitimate explanatory questions. A legitimate question takes the form of "why *P* rather than *Q*?" and presupposes the knowledge of *P* and not-*Q*. But as with the ontic view, the epistemic view adds that this question selects for certain kinds of explanatorily relevant factors in the world. A genuine explanation will then be an account that picks out some facts that do bear the right kind of relation to the contrastive question. The ontic view claims that all the features of a genuine explanation relate only to facts in the world, and that the characteristics of agents are irrelevant. The epistemic view maintains that the world plays an important role, but that a full account of what makes an explanation genuine must start with legitimate explanatory questions. Which questions are legitimate will vary with a person's states, especially their states of knowledge and interests. However, on this epistemic view, that is the only role for context and pragmatics.

Each strategy faces its challenges. The ontic view must clarify the nature of contrastive facts and their relationship to non-contrastive facts. The epistemic view needs to flesh out what makes a question legitimate. If all questions are legitimate, then we risk trivializing explanation (Kitcher and Salmon 1987). Either way, the arguments of this chapter show that it is not easy to make sense of explanatory pluralism. Whatever strategy turns out to be the best, there are many extant approaches to explanation that fail to accommodate explanatory pluralism while doing justice to the value that scientists place on discovering genuine explanations.

Acknowledgments

I am grateful to the editors and several anonymous referees for their helpful comments on an earlier draft of this chapter.

References

- Andersen, H. (forthcoming), 'Complements, Not Competitors: Causal and Mathematical Explanations', *British Journal for the Philosophy of Science*.
- Baron, S., Colyvan, M., and Ripley, D. (forthcoming), 'How Mathematics Can Make a Difference', *Philosophers' Imprint*.
- Brigandt, I. (2013), 'Explanation in Biology: Reduction, Pluralism, and Explanatory Aims', *Science and Education* 22: 69–91.
- Garfinkel, A. (1981), *Forms of Explanation: Rethinking Questions in Social Theory* (New Haven: Yale University Press).
- Haslanger, S. (2016), 'What Is a (Social) Structural Explanation?', *Philosophical Studies* 173: 113–30.
- Hempel, C. (1965), *Aspects of Scientific Explanation* (New York: Free Press).
- Hitchcock, C. (2012), 'Contrastive Explanation', in M. Blaauw (ed.), *Contrastivism in Philosophy* (New York: Routledge), 11–34.
- Kitcher, P. and Salmon, W. (1987), 'Van Fraassen on Explanation', *Journal of Philosophy* 84: 315–30.
- Koslicki, K. (2012), 'Varieties of Ontological Dependence', in F. Correia and B. Schneider (eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (Cambridge: Cambridge University Press), 186–213.
- Lipton, P. (2004), *Inference to the Best Explanation*, 2nd edn. (New York: Routledge).
- Lipton, P. (2008), 'CP Laws, Reduction and Explanatory Pluralism', in J. Hohwy and J. Kallerstrup (eds.), *Being Reduced: New Essays on Reduction, Explanation and Causation* (Oxford: Oxford University Press), 115–25.
- Pincock, C. (2007), 'A Role for Mathematics in the Physical Sciences', *Noûs* 41: 253–75.
- Pincock, C. (2015), 'Abstract Explanations in Science', *British Journal for the Philosophy of Science* 66: 857–82.
- Potochnik, A. (2015), 'The Diverse Aims of Science', *Studies in History and Philosophy of Science* 53: 71–80.
- Reutlinger, A. (2016), 'Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation', *Philosophy of Science* 83: 733–45.
- Rice, C. (2015), 'Moving Beyond Causes: Optimality Models and Scientific Explanation', *Noûs* 49: 589–615.
- Saatsi, J. (2016), 'On the "Indispensable Explanatory Role" of Mathematics', *Mind* 125: 1045–70.
- Saatsi, J. and Pexton, M. (2013), 'Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations', *Philosophy of Science* 80: 613–24.
- Shapiro, L. and Sober, E. (2007), 'Epiphenomenalism: The Do's and Don'ts', in G. Wolters and P. Machamer (eds.), *Thinking About Causes: From Greek Philosophy to Modern Physics* (Pittsburgh, PA: University of Pittsburgh Press), 235–64.
- Sober, E. (1986), 'Explanatory Presupposition', *Australasian Journal of Philosophy* 64: 143–9.
- Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation* (New York: Oxford University Press).
- Woodward, J. (2015), 'Interventionism and Causal Exclusion', *Philosophy and Phenomenological Research* 91: 303–47.
- Woody, A. (2015), 'Re-orienting Discussions of Scientific Explanation: A Functional Perspective', *Studies in History and Philosophy of Science* 52: 79–87.
- Wright, C. (2015), 'The Ontic Conception of Scientific Explanation', *Studies in the History and Philosophy of Science* 54: 20–30.

3

Eight Other Questions about Explanation

Angela Potochnik

1. Introduction

Philosophical accounts of scientific explanation are by and large categorized as law-based, unificationist, causal, mechanistic, etc. This type of categorization emphasizes one particular element of explanatory practices, namely, the type of dependence that is supposed to do the explaining. This question about scientific explanations is: in order for *A* to explain *B*, in what way must *A* account for *B*? Various philosophers have answered this question with the suggestion that, to explain, *A* must account for *B* according to natural law, or by reduction to an accepted phenomenon, or in virtue of causal dependence, or by mechanistic production, etc. Accordingly, students of philosophy of science are introduced to the deductive-nomological account, the unification account, various causal accounts, the mechanistic account, etc.¹ In recent years, causal accounts and mechanistic accounts, which also require causal dependence, have enjoyed broad appeal.

There are, of course, many other features of explanatory practices aside from the type of dependence that counts as explanatory. And philosophers disagree significantly about the nature of some of these other features as well. But those disagreements tend to be formulated as downstream issues about a particular account of explanation. In other words, the defining feature of an account of explanation is typically the posited form of explanatory dependence—is it a causal account, a law-based account,

¹ This categorization is of course not exhaustive, and it conceals a great deal of variety, for instance in how causes are to be understood for a causal account of explanation. What is important for present purposes is simply the element of explanatory practices that such a categorization focuses upon, namely, what form of dependence is explanatory. This construal is more commonly attached to causal and mechanistic accounts of explanation than to unification or D-N accounts, but I believe it suits the latter accounts as well. Friedman (1974), a prominent advocate of a unification account, articulates the question of explanation as that of the relation between the phenomenon explained and the phenomenon doing the explaining. The D-N requirement of citing a natural law also coheres with this construal; that amounts to the requirement that *A* account for *B* in virtue of natural law.

or something else? Only once this is settled do most philosophers consider other elements of explanatory practices. For example, one might embrace Woodward's version of a causal account of explanation, where causation is understood in terms of difference-making and invariance is taken to be explanatorily important. This leads to an emphasis on the value of general explanations like the ideal gas law (see Woodward 2003). Or one may embrace Salmon's version of a causal account of explanation, where causation requires mark-transmission and the explanatory value of causal processes is taken to be central (see Salmon 1984). This disqualifies some of the explanations that Woodward emphasizes, including the ideal gas law (or, at least, that is Salmon's view). In light of the prevailing philosophical focus on the type of explanatory dependence, though, these deep disagreements are treated as ancillary concerns that merely distinguish different varieties of the causal account of explanation.

Overemphasis of this single element of explanatory practices has, I believe, eclipsed the significance of several other features of scientific explanations and philosophical disagreements about those features. In this chapter I articulate eight such features and some of the philosophical views about each. I note dependencies among views of different features of explanation where those exist. But by and large, these are eight distinct and independent questions that can be posed about the nature of scientific explanation—or nine questions, if we include the question about the explanatory dependence relation(s). The purpose of this is not to develop an account of explanation nor to defend any one conception of these features. Instead, the aim is to further philosophical debate about the nature of scientific explanation by distinguishing among relatively independent features of explanatory practices and, for each, clarifying what is at issue. These various features of explanation fall roughly into three categories, reflected in the following three sections. There are questions to be asked about the role of human explainers in the project of scientific explanation (section 2); representational questions about what explanations should actually be formulated and the relationship those explanations bear to other scientific projects (section 3); and finally, ontological questions surrounding what, out in the world, explains (section 4). This last category includes the classic question of what form of dependence is explanatory, but it includes other questions as well.

Philosophical progress does not always involve resolving the main dispute. My aim here is to contribute to a different kind of progress, namely, drawing attention to philosophical questions about scientific explanation that are distinct from whether all explanations require citing causal dependences and other questions about the nature of explanatory dependence. It is in that sense that this chapter is about explanation beyond causation. I hope this results in the identification of features of explanation that have not been sufficiently explored, clarification of what is at stake between opposed views about those features, and thus the development of a more nuanced understanding of the philosophical issues surrounding scientific explanation. I believe there are at least eight questions to ask about scientific explanation, aside from whether causal dependence relations are always or ever explanatory. Let us now consider them.

2. Human Explainers

I begin by exploring open issues regarding human explainers. This may seem odd, given the overwhelming emphasis in the literature on the explanatory dependence relation, a question about ontology. But, as will become clear further below, I do so for a principled reason. There are two kinds of questions about human explainers. First, one can ask how the people doing the explaining, and the audiences for those explanations, influence explanatory practices. Second, one can ask to what degree those influences are relevant to a full-fledged account of explanation. I will begin with the latter question, whether philosophical accounts of explanation should address human influences on explanatory practices.

Question 1: Priority of communication

A debate has recently emerged, or perhaps been revived, surrounding the so-called ontic versus communicative senses of explanation. This is at root a debate about the significance, or lack thereof, of human explainers to a philosophical account of scientific explanation. Proponents of an ontic or ontological approach to explanation judge the important features of scientific explanation to be independent of human influences. This includes independence from who in particular is doing the explaining, as well as the fact that all explanations are formulated by humans. A position like this has been advocated at different times and in different contexts by David Lewis (1986), Wesley Salmon (1989), Michael Strevens (2008), and Carl Craver (2014), among others. Other philosophers have adopted the opposed view that human explanatory practices must be the starting point for any account of explanation. Notable instances of this view include Sylvain Bromberger's (1966) treatment of why-questions, Bas van Fraassen's (1980) pragmatic account of explanation, and Peter Achinstein's (1983) illocutionary account. In contrast to a primarily ontic or ontological approach, one might think of these views collectively as a communicative approach to explanation. They all focus substantially on the communicative roles explanations are formulated to play, and look there for insight into the nature of scientific explanation. I have also motivated a communicative approach to explanation (see Potochnik 2015a, 2016). Ontic and communicative approaches thus provide two different answers to the question about the priority of communication to an account of explanation: the former judges the specificities of human explainers to be irrelevant to a philosophical account of explanation, the latter takes them to be central.

One role of human explainers is wholly uncontroversial. Humans, and particular individuals at that, are responsible for formulating the requests for explanation. This means that human characteristics and idiosyncrasies find their way into what *explananda* are targeted by scientific explanations—that is, what events scientists attempt to explain and how those events are characterized. Some think this influence extends also to a more fine-grained characterization including not only the event to be explained, but also the alternative state of affairs the event is to be contrasted

with, often referred to as the explanandum's contrast class. According to a contrastive approach to explanation, different explanations are warranted when explaining why a car crashed at night rather than not crashing at all, versus why a car crashed at night rather than crashing during the day.

From an ontic perspective, once the explanatory agenda is set (the explanandum specified, and perhaps the contrast class as well), the proper human influence on scientific explanations has been exhausted. All the remaining work is done by an account of explanatory dependence. The explanatory agenda simply determines what, out in the world, explains a given event. From a communicative perspective, in contrast, this is just the tip of the iceberg. Human influences on scientific explanations are taken to extend beyond setting the explanatory agenda, in one way or another influencing which explanation satisfactorily accounts for some explanandum and contrast class. For example, on van Fraassen's (1980) account, human characteristics and concerns also influence the explanatory relation itself, that is, the relationship an explanation should bear to the event to be explained.

If human explainers, their interests and idiosyncrasies, are taken to be central to the enterprise of explaining, then other questions are raised about the relationship an explanation must bear to its audience, and what is required for an explanation to succeed in explaining. For this reason, much of what I say below about the other questions about human explainers presupposes a communicative approach to explanation. One can certainly recognize additional questions about human explainers without adopting a communicative approach to explanation. It's just that, from an ontic perspective on explanation, these further questions will tend to be seen as unimportant to philosophical questions about scientific explanation. For instance, Lewis (1986) dismisses questions around the "pragmatics" governing explanation as not distinctive questions for scientific explanation, but questions about human discourse in general. Similarly, a proponent of an ontic approach may take there to be interesting questions about the psychology of explanation, but deem these incidental to a philosophical account of explanation.

Question 2: Connection to understanding

Another question about the human element of explanation that has recently received more attention is the nature of the relationship between explanation and understanding. The basic question is whether explanation and understanding are inextricably linked. One might wonder whether any explanation must result in understanding in order to succeed. And one might wonder whether any and all understanding must issue from an explanation.²

Consider, first, the question of whether an explanation is necessary for understanding. Peter Lipton (2009) has argued that understanding can be possessed in circumstances in which we would hesitate to say there is an explanation. One such circumstance is

² De Regt (2013) provides a nice summary of the debate surrounding these questions.

understanding via tacit causal knowledge gained from images, the use of physical models, or physical manipulations. Lipton also argues that understanding can emerge from examining exemplars, or from modal information. In his view, none of these sources of understanding are of the right sort to give rise to *explanations* of the phenomena they help one understand. This is because, according to Lipton, an explanation must be able to be communicated, at least to oneself (so cannot be tacit), and must contain information about the object of understanding, that is, about why something in fact came about (which modal information arguably does not). Notice that the first of these requirements presumes something about the human element of explanation, namely, that any scientific explanation must play the proper communicative role.

Strevens (2013), in contrast, argues that there is no understanding but by way of explanations. In his view, understanding a phenomenon just is to grasp a correct explanation of that phenomenon. Strevens responds directly to some of Lipton's purported cases of understanding without explanation. He disputes Lipton's claim that explanations must be explicit, able to be communicated; in his view, tacit understanding simply arises from grasping a tacit explanation. Strevens and Lipton thus disagree about a prior issue, namely the significance of the communicative sense of explanation. As we have already seen, Strevens adopts an ontic approach, deeming the communicative purposes of explanations unimportant to an account of explanation. Strevens also argues that, when something tacit like physical intuition is the source of understanding, this understanding arises only in virtue of the *accuracy* of the physical intuition. He says, of a particular example, "it amounts to genuine understanding why, I suggest, only insofar as the psychologically operative pretheoretical physical principles constitute a part of the correct physical explanation" (Strevens 2013: 514). For Strevens, it is precisely the ontic element of explanations—that they track an explanatory dependence relation—that is supposed to fill the gap between intuition and legitimate explanation.

Besides this debate of whether explanation is necessary to generate understanding, there is also a question of whether any explanation must be sufficient to produce understanding. Can there be a (successful) explanation that does not generate understanding, or that does not even have the potential to do so? This question seems to not often be addressed explicitly, at least not as formulated here. But a position on the issue is suggested by those who affirm the importance of an account of explanation also accounting for the production of understanding. This move is one way of affirming the importance of an explanation connecting in the right way to its human audience. For example, Hempel (1965) motivated the classic deductive-nomological account of explanation with the idea that deductions from laws of nature show that "the occurrence of the phenomenon was to be expected", and that "it is in this sense that the explanation enables us to understand why the phenomenon occurred" (337). Explanatory dependence relations out in the world are clearly insufficient for producing understanding. To generate understanding, information about those relations must be communicated to an

audience, and must be communicated in a way that leads to the cognitive achievement of understanding. The opposite view on this question—that explanations need not generate understanding—seems to follow from a strongly ontic approach to explanation, where explanations exist out in the world, even if they are never identified or communicated.

Question 3: Psychology of explanation

A third topic that relates to human explainers is the psychology of explanation. Explanation in general and scientific explanation in particular is a topic of empirical research in cognitive psychology. That research aims to uncover the cognitive roles played by explanation, and what features accepted explanations tend to possess. For example, Tania Lombrozo (2011) surveys empirical research that suggests the act of explaining improves learning of general patterns and causal structure. She also discusses research suggesting a broad preference for simple explanations and explanations that are highly general. Philosophical accounts of explanation can differ in the degree of importance they attach to the psychological elements of explanation, the type of relevance those psychological elements are supposed to have, and (if relevant) which psychological elements of explanation they take to be significant.

If the communicative roles explanations play are taken to be central to the nature of explanation, then why and how explanations are in fact formulated is directly relevant to a philosophical account of explanation. On this approach explanations cannot succeed without being accepted as explanatory, so what features humans value in explanations and explanations' cognitive purposes influence the features explanations should possess. Some advocates of a strongly ontic approach to explanation instead hold that the important features of explanation are independent of the features of those formulating and receiving explanations. In that case, research into the psychology of explanation is at most indirectly relevant to the norms of explanation. Our intuitions about what is explanatory may track the norms of explanation, but they cannot influence them.

3. Explanations as Representations

A second category of philosophical questions about scientific explanation regards representation. As with human explainers, one can ask what relevance representational decisions have to a philosophical account of scientific explanation. And, as with the first category of questions, granting a role for questions of representation introduces downstream questions, such as what should be represented in an explanation, and with what fidelity. These are questions about the role that abstraction and idealization should play in scientific explanations. Finally, as I discuss below, debate about the representational features of explanation relates also to questions about the relationship between explanation and other scientific aims.

Question 4: Priority of representation

Just as one can question whether human explainers and explanations' communicative and cognitive roles shape scientific explanations in a philosophically significant way, so too one can ask whether representational decisions shape scientific explanations in a way that is central to providing a philosophical account of explanation. Since representational decisions can be made for purposes of improved communication or cognition, these two questions may be related, and I suspect they have sometimes been conflated. But some who embrace an ontological approach to explanation afford a central role in an account of explanation to representational decisions, but not for communicative or cognitive purposes. A prime example is Strevens's (2008) kairetic account of explanation. Strevens develops what he calls a two-factor account of explanation. The first factor is an account of the type of metaphysical dependence relation that can be explanatory, and the second factor is a separate account that determines which facts about such relations belong in any given explanation. This second factor is at least in part a question of representation. Evidence of this is that a central feature of Strevens's account is the determination of the right degree of generality, or abstractness, of an explanation. This is a matter about how to represent the world—with greater or less detail. Indeed, in Strevens's view, citing a general law simply is to cite the underlying physical mechanism, but the former is a better explanation (see Strevens 2008: 129–30). The difference can't be metaphysical, then, but representational.

And so, within an ontological (versus communicative) approach to explanation, there is still a question of primacy to an account of explanation of facts out in the world or how we go about representing those facts. Some proponents of an ontological approach think that the ontological side—the nature of explanatory dependence relations—is where all of the work, or at least all of the important work, is located. For a good example of this, see Craver (2014). Others, like Strevens, think there are significant questions about how the explanatory dependence relations are represented.

Also analogous to, but distinct from, the case of the ontological/communicative divide is the question of whether the ontological dimension of explanation is always “upstream” from, that is logically prior to, any representational dimension of explanation. This can be understood as the question of what needs to be settled first in order to get traction on any other questions about explanation. On this I believe Strevens and Craver would agree: the type of explanatory dependence, and the nature of that dependence in some particular phenomenon to be explained, must be settled first. Put another way, their view is that making true claims about explanatory dependence is the primary determinant of the content of explanations. Arnon Levy (n.d.) suggests, against this kind of a view, that the “goodness” of an explanation might be enhanced by sacrificing some truth. This might be so if explanations can be improved by incorporating idealizations, or assumptions recognized as false.³ One such view is advocated

³ Strevens (2008) has a view of idealizations' explanatory role that does not stray in this way from a fully ontological approach to explanation.

by Robert Batterman (see, e.g., Batterman 2002, 2009). He argues that one central form of explanation, what he calls asymptotic explanation, is impossible without idealization. If this is right, it requires granting that some questions about *how* our explanations should represent must be settled prior to—or at least independently from—what, out in the world, they should represent.

Question 5: The representational aims of explanation

The weaker claim articulated above about the representational features of explanations is that those features can be distinctive and warrant consideration, even if they are “downstream” from explanations’ ontological features. If one grants at least this much, then this introduces questions about what, and how, the explanations generated in science should represent. In particular, when (if ever) should explanations represent more abstractly, by including less detail, and when (if ever) should explanations represent less accurately, by including idealizations? If one holds the stronger view that the representational requirements for explanation can influence explanations’ ontological features, then this opens up additional possibilities for when explanations should omit or falsify some details. Views abound about the role of abstraction and idealization in scientific explanations; some of those views suggest this weaker commitment regarding the representational features of explanation, whereas others require the stronger.

Consider first the matter of an explanation’s abstractness. Is more detail (about explanatorily relevant dependence) always better than less detail? Or are explanations ever improved by omitting information? The issue is a bit subtle, as much rides on what is built into the determination of “explanatorily relevant dependence”. This is an ontological issue, and as such, I’m postponing it until section 4. Returning to Strevens’s view provides an illustration of both the subtlety and also a position on the question of abstraction. At first glance, Strevens’s answer is, definitively, that explanations should leave out lots of information. For him, the raw material of explanations is causal entailment; this is the first factor in his two-factor account. But then there’s a question of which representations of causal entailment are most explanatory; answering this is the job of the second factor. Strevens argues that only causal factors that are difference-makers (in his sense) should be included in an explanation; this results in explanations with the right degree of generality and abstractness.

But this doesn’t fully settle the issue for Strevens, as there’s still a question of how many difference-making factors an explanation should feature. Should explanations be “elongated”, that is, expanded to include factors that made a difference to the cited difference-making factors? Should explanations be “deepened”, that is, expanded to include a physical explanation for any high-level laws that are cited? Both of these are ways of incorporating additional details and, thus, making explanations less abstract, but they are distinct issues from each other, and distinct also from the first way in which Strevens thinks explanations should be abstract. Strevens’s answers are that elongation is optional but it improves an explanation, and that deepening is compulsory (see, e.g., 2008: 133). However, this is not so for “causal covering-laws”, such as the

kinetic theory of gases, since as I mentioned above, Strevens thinks that citing such a law is the same thing as citing the underlying physical mechanism (2008: 129–30).

I said that Strevens's view illustrates not only how one might take abstractness to be a desirable feature of explanations, but also the subtlety of the issue. Strevens encourages abstract explanations in one sense (omitting non-difference-makers), while allowing them and prohibiting them in two other senses (non-elongated explanations and non-deep explanations, respectively). As for the subtlety of the issue, it is difficult to determine which of these positions concerns the question of what things are explanatory (i.e., the ontological element of explanation) and which, if any, concerns the question of how explanatory things should be represented. That non-difference-makers should always be omitted seems to be an ontological question of what facts about the world are explanatory; Strevens holds that only difference-makers (in his sense) explain. Yet the matter is murkier for his positions regarding elongation and depth. Elongation seems to be a question of how many of the explanatory dependence relations to represent, so perhaps this issue is not ontological but representational. I find the requirement of depth to be more puzzling still. Strevens claims that this requirement is "quite consistent with a high degree of abstraction" (2008: 130), and that an abstract causal covering-law is, from an ontological perspective, one and the same explanation as the physical mechanism(s) underpinning it. He says the former has a "communicative shortcoming" but not an "explanatory shortcoming" (131). But this suggests that determination of difference-making is, for Strevens, not purely an ontological matter after all. A causal covering-law omits information about the underlying physical mechanism because those details are not difference-makers. But the ontological explanation provided by a causal covering-law is supposed to be the same as what would be provided by citing the underlying physical mechanism. The determination of difference-making seems, then, to regard not the ontological explanation but what details are included—that is, *represented*—in a causal model.

There are, of course, other views about how abstract explanations should be. Like Strevens's, these other views are by and large developed within the structure of particular accounts of the explanatory dependence relation. But it needn't be so. One might bracket the issue of the nature of explanatory dependence by approaching the issue of explanations' abstractness from the perspective of existing explanatory practices and findings about explanation from cognitive psychology (introduced as Question 3 above).

Let's move on to the issue of explanations' fidelity, that is, whether explanations can and should include idealizations. As I mentioned above, one notable advocate of idealized explanations is Batterman (2002, 2009). Batterman argues that there is an important style of explanation, what he calls asymptotic explanation, that relies essentially on the use of idealizations. Roughly, the idea is that explanations of how phenomena behave as they approach a limit are enabled by idealizing parameters as having an extreme value of zero or infinity. If this is right, some explanations are impossible without including idealizations. In contrast, John Norton (2012) acknowledges the

importance of this style of explanation, but he disputes the claim that setting a parameter to zero or infinity is an idealization; he takes these simply to be approximations. Like Batterman, Strevens also defends the explanatory value of idealizations, but he limits their role to standing in for non-difference-makers, thereby expressing what did *not* make a difference to the phenomenon. Alisa Bokulich (2011) endorses a position somewhat between these views, for she argues that “fictionalized” representations can explain, but that they do so by correctly capturing the explanatory counterfactual dependence. It’s worth pointing out that Bokulich takes such explanations to be non-causal in virtue of the fictions they incorporate, because in her view fictional entities cannot have causal powers. This is a view about the ontological question of explanatory dependence that is informed by a position regarding the representational question of idealized explanations, rather than the other way around.

Many other philosophers have views about idealizations’ role in explanation, but I will mention my own view as a final example, since I take it to contrast nicely with Strevens’s and to exemplify a view of the relationship between communicative, representational, and ontological elements of explanation opposed to his. I think explanations employ idealizations not only to signal what did not make a difference to the phenomenon, but also (and much more commonly) simply to signal that researchers’ interests lie elsewhere (Potochnik, 2017). Adopting for the nonce Strevens’s view of the explanatory dependence relation, even important difference-makers might be idealized away in order to simplify an explanation and draw attention to other difference-makers, the ones in which those formulating the explanation are primarily interested. This reverses the priority of communicative and ontological features of explanation. In my view it is the communicative or psychological needs of an explanation’s audience that determines what should be veridically represented and what should be omitted or falsified, and that determination in turn sheds light on what sort of dependence is explanatory. I will not defend this idea here; I simply mention it as an alternative view of the explanatory role of idealizations.

Question 6: Relationship to other scientific aims

Another question about scientific explanation regards its role in the scientific enterprise. In particular, one might wonder how explanation relates to other scientific aims. For example, Heather Douglas (2009) argues that the role of explanation in generating good predictions has been overlooked, and that this has weakened accounts of explanation. She says that explanations are a cognitive tool to aid in generating predictions, for they “help us to organize the complex world we encounter, making it cognitively manageable” (54). In direct opposition to this idea, I have argued that different scientific aims, including explanation and prediction, motivate different types of scientific activities and products (see Potochnik 2010a, 2015b, 2017). On this view, a perfectly good explanation, such as an explanation that idealizes many important causal influences in order to represent the causal role of just one kind of factor, may be poorly suited as the basis for making predictions.

One might wonder why I include this in a list of questions about representational features of explanation. For one thing, notice that the two views I briefly characterized both regard explanations in their representational sense. Douglas's description of explanations as cognitive tools clearly is not about what facts out in the world are explanatory, but the useful ways in which scientists represent those explanatory facts. Only facts that are known and represented can be cognitive tools. Similarly, my contrasting view is not a view about the ontological dimension of explanation: whatever dependencies are explanatory presumably are also helpful in the formulation of predictions. The question is whether explanations actually formulated should also lend themselves to generating accurate predictions. A view on this issue will have implications for the kind of representations our explanations should be, including their abstractness and fidelity. If explanations should support accurate predictions, then they must be accurate enough, and specific enough, about the full range of the applicable dependence relations to play this role. A strong view of the explanatory role of idealization thus commits me to a division between explanation and other scientific aims, including prediction.

4. Ontic Explanations

The third category of philosophical questions about scientific explanation I will discuss regards ontology. As with human explainers and the representational form of explanations, the two categories of questions discussed above, there is a question of how central the ontological dimension of explanatory practices is to a philosophical account of explanation. There are also questions about the nature of this ontological dimension, that is, the form(s) of explanatory dependence. In contrast to the issues I have surveyed surrounding human explainers and representation, few in any deny that explanations' ontological dimension is central to providing a philosophical account of explanation. Accordingly, almost all philosophers who address scientific explanation engage with one or another ontological question about explanation, or at least grant the significance of those questions. Indeed, I suggested at the outset of this chapter that attention to the nature of the explanatory dependence relation, which I take to be an ontological question, tends to eclipse many of these other disagreements about explanation. I begin the present section by discussing this question that's at the center of so many philosophical accounts of explanation. I then move on to the question of the priority of the ontological dimension of explanation, and then discuss a further, arguably ontological question about explanation, namely the issue of level(s) of explanation.

The question of the nature of explanatory dependence

I have suggested that one ontological issue about explanation gets an undue share of philosophical attention. This is the matter of the explanatory dependence relation,

the question of what, out in the world, explains.⁴ Many a philosophy of science course has contained a unit on scientific explanation that looks something like: scientific laws explain!; no, it must be causes; but, unification! This perhaps is continued with: causal mechanisms explain; or is it causal difference-makers? The more general question is sometimes introduced of whether there's a unitary account to give of the form of explanatory dependence. This is often yoked to the question of whether purely mathematical dependencies can ever be explanatory.

This question of what form(s) of dependence are of explanatory value in science is undoubtedly important, and the debate about how to answer this question rages on. Versions of a causal account of explanation have dominated the literature in recent decades, which is part of the motivation for this volume's focus on non-causal explanation. Above I described how Bokulich rejects a causal approach to causation because of the extensive fictions employed in explanations. Others who have challenged a causal approach focus directly on the nature of explanatory dependence. Some who have emphasized the explanatoriness of broad patterns think this undermines the idea that explanatory dependence is always causal. This includes, notably, advocates of the unification approach (see Friedman 1974), but also Batterman (2002) and others. Some of these accounts share with Bokulich's an acceptance of the explanatory significance of difference-making, while denying that difference-making constitutes causal influence. Others focus on cases when the explanatory dependence seems to be purely mathematical (see Pincock 2012; Lange 2013).

This is an important, live debate. But I hope it is clear from what I have said so far that developing a view of the explanatory dependence relation is not in itself sufficient to provide a philosophical account of scientific explanation. Too many other questions are left unanswered. Of course, many proponents of one or another view about the explanatory dependence relation have much to say about some of these other issues surrounding explanation. But far too often, those other issues are treated as merely add-on features to a core account, an account that is named for its commitment to some form of explanatory dependence. Instead, they are separate, partially independent questions about the nature of scientific explanation.

Question 7: Priority of the ontological dimension

I suspect that one reason the nature of the explanatory dependence relation has received the lion's share of philosophical attention is the common presumption that the ontological dimension of explanation is primary, or even solitary, in its importance. This raises the next question about the ontology of explanations I want to discuss, namely the centrality of this dimension as compared to the representational and communicative dimensions of explanation. This is the counterpart of Questions 1 and 4

⁴ Note that accounts of explanatory dependence vary in the degree to which they are strictly ontic. For example, the deductive-nomological account takes explanation to occur among propositions about phenomena and laws, whereas Craver (2014) argues that explanations are ultimately relations among phenomena out in the world.

in the previous two sections, about the priority of communication and representation, respectively, for explanation.

Few deny that dependence relations out in the world are relevant to what qualifies as an explanation. For our scientific explanations to succeed, they must track some dependence—of the right kind—that actually exists in the world. Perhaps van Fraassen (1980) comes the closest to denying this, since he argues that there is not a unitary account to be given of explanatory dependence relations, that this depends on an explanation's communicative context. As we have already seen, many others think that the ontological issue of explanatory dependence is where all the work in providing an account of explanation, or at least all the important work, is located. Communicative influences are often relegated to the category of the “pragmatics” of explanation, and Lewis (1986) influentially argued that the pragmatics of explanation is nothing special, that is, is in no way distinct from the pragmatics of linguistic communication more generally. Craver (2014) holds an extreme version of an ontological, or ontic, view of explanation. He argues that what counts as an explanation is purely an ontological matter, not representational or communicative, for “our abstract and idealized representations count as conveying explanatory information in virtue of the fact that they represent certain kinds of ontic structures (and not others)” (29).

Views about the priority of the communicative sense of explanation or representational issues in explanation, the first and fourth questions discussed above, have obvious implications for this issue. If one grants the significance, or even primacy, of the audience's influence on the content of an explanation, then this amounts to rejecting a purely ontological approach to explanation. And if one grants the importance of representational matters, including whether and how explanations should abstract and idealize what they represent about the world, then one has at least strayed from an extreme ontic view like Craver's. In contrast, a commitment to a view like Craver's or Lewis's can—and has—been used to justify producing an account of explanation that consists solely of a view about the nature of explanatory dependence. Other views are in a confusing middle ground. As we saw in section 3, Strevens explicitly claims that his account of explanation is ontological in nature, yet a good deal of that account focuses on representational issues, including both abstraction and idealization.

Question 8: Level of explanation

Another well-identified question about explanation regards the proper *level* of explanation. Unlike many of the other questions about explanation I've surveyed so far, this issue is often treated separately from providing an overarching account of explanation. It also has been linked to positions on a range of other issues in philosophy of science, for example, about reductionism, ontology, and the relationships among different fields of science. Classic, reductionist approaches to the unity of science claimed that the reduction of all scientific findings to microphysical laws and happenings entailed the successful explanation of those findings in microphysical terms (see, e.g., Hempel and Oppenheim 1948). An opposed position is to declare that some explanations are

benefited from being at a higher level than microphysics. This idea has been developed in a variety of ways by different philosophers over the years. In this context, “higher level” might mean more abstract, more general, invoking bigger entities, invoking laws outside of microphysics, or some combination of these. Putnam (1975) memorably illustrated high-level explanation with the example of explaining why a square peg with one-inch sides did not fit through a round hole with a one-inch diameter. There continue to be proponents of high-level explanation (see, e.g., Weslake 2010), pluralism about the proper levels of explanation (see, e.g., Potochnik 2010b), and explanatory reductionism (see, e.g., Kim 2008).

The question of the proper level of explanation is plausibly about the ontological dimension of explanation. One might phrase the question as: what are the kinds of things that can explain? Are these always only microscopic particles and the laws governing them, or sometimes middle-sized objects and the relationships among them? And examples of these options are, respectively, the molecular structure of Putnam’s peg and board, and the geometric relationship obtaining between the peg and the hole in the board and the rigidity of the two objects. On the other hand, one might think of the question of the proper level of explanation as primarily or solely regarding representational decisions. Recall Strevens’s claim that to cite a causal covering-law just is to cite the physical mechanism responsible for said law. It seems that, in his view, the ontological element of those explanations is identical—all that distinguishes them is representational differences. Yet one of the two explanations is at a higher level, in the sense of being more abstract and avoiding reference to the fundamental physics of the phenomenon. I’m not inclined to accept this interpretation of the issue. I agree, of course, that the proper degree of abstraction is a representational issue. But in my view, representational decisions can’t help but influence explanations’ ontic features, that is, what out in the world explains (see Potochnik 2016).

5. Conclusion

I began this chapter with the suggestion that the debate about the nature of explanatory dependence has eclipsed several other philosophical questions about scientific explanation. What followed, in the bulk of the chapter, was a rapid-fire listing of eight of these other questions, with brief discussions of the nature of each question and a sampling of views about them. I have tried to articulate these questions about explanation in a way that clarifies any relations of dependence among views about different questions, and that emphasizes the independence of each from an account of the explanatory dependence relation.

These questions about explanation fall, roughly, into three categories. They are: questions about the human element of explanation, that is, whether and how explanations are shaped by communicative purposes and cognitive needs (section 2); questions about the representational element of explanation, that is, whether and how explanations are shaped by representational decisions (section 3); and questions about the ontic element

of explanation, that is, how explanations are shaped by features of the world and the relationships they bear to the phenomena to be explained (section 4). The logically primary question in each category is whether and to what degree that element of explanation is relevant to giving a philosophical account of explanation. Other questions in each category regard the nature of that element's relevance. For the human element of explanation, these questions include how explanations (generated by humans) relate to human understanding, and the cognitive psychology of explanation. For the representational element of explanation, these questions include how explanations should represent—in particular whether and when they should abstract and idealize, and the relationship explanations generated in science bear to other scientific aims, such as prediction. Finally, for the ontic element of explanation, there's the familiar question of the nature of explanatory dependence, as well as the question of the proper level(s) of explanation.

Historically, the ontic element of explanation has been presumed to be of either central or sole relevance. Even accounts of explanation that focus on explanations in the representational sense, such as the deductive-nomological and unification accounts, have placed the source of explanatoriness on the ontic side—e.g. for the D-N account, the laws of nature cited and facts accurately described, and for Friedman's (1974) unification account, in a relation among phenomena. With a few prominent exceptions, there has been little attention devoted to defending the centrality of the ontic element of explanation. In contrast, attention to communicative elements of explanation must always begin with a defense of the relevance of those issues, or else risk the dismissive response that the discussion is irrelevant to the real issues about explanation. I began this chapter with questions about the human element of explanation in order to demonstrate that the traditional ordering of priorities for an account of explanation is not inevitable. Despite the strong precedent for accounts of explanation that are ontic-first or ontic-only, there are significant questions about how our explanations are shaped by communicative purposes and cognitive needs, and whether and how these are distinctively human. Those questions often can be addressed directly, rather than merely as add-on components to an account of the ontic element of explanation. Furthermore, how these questions about the communicative element of explanation are answered can have implications for an account of the ontic element of explanation. This is so for my own view of explanation (see Potochnik 2017).

The recognition that there are other questions about explanation is, of course, not uniquely mine. As I have surveyed here, there already exists philosophical work on most or all of the topics I've listed. My hope is that the contribution of this chapter consists partly in the delineation and categorization of these many issues, and partly in the demonstration of their distance from the question of what, out in the world, explains. My aim in surveying so many questions is to illustrate the vast space for different kinds of disagreements about scientific explanation. Surely other philosophical questions about scientific explanation exist even beyond those I have detailed here. Philosophers of science working on, or considering work on, the nature of scientific

explanation: I urge you to consider this range of largely independent questions about scientific explanation. Choose a question to explicitly develop a view on; show inter-relationships among views one might hold about a few of these features; articulate still further questions in need of answers. If you must, develop a new account of the sort of dependence that is explanatory. But please, do not be convinced that the main philosophical question about explanation is whether causes, laws, or something else are the kind of thing that explains.

Acknowledgments

Thanks to the editors for including me in this project and for their effective leadership of the project. The ideas and prose of this chapter were significantly improved by a reviewer for this volume.

References

- Achinstein, P. (1983), *The Nature of Explanation* (Oxford: Oxford University Press).
- Batterman, R. W. (2002), *The Devil in the Details* (New York: Oxford University Press).
- Batterman, R. W. (2009), 'Idealization and Modeling', *Synthese* 169: 427–46.
- Bokulich, A. (2011), 'How Scientific Models Can Explain', *Synthese* 180: 33–45.
- Bromberger, S. (1966), 'Why-Questions', in R. Colodny (ed.), *Mind and Cosmos* (Pittsburgh: University of Pittsburgh Press), 86–111.
- Craver, C. F. (2014), 'The Ontic Account of Scientific Explanation', in M. I. Kaiser, O. R. Scholz, D. Plenge, and A. Hüttemann (eds.), *Explanation in the Special Sciences: The Case of Biology and History* (Dordrecht: Springer), 27–52.
- de Regt, H. W. (2013), 'Understanding and Explanation: Living Apart Together?', *Studies in History and Philosophy of Science* 44: 505–9.
- Douglas, H. (2009), 'Reintroducing Prediction to Explanation', *Philosophy of Science* 76: 444–63.
- Friedman, M. (1974), 'Explanation and Scientific Understanding', *Journal of Philosophy* 71: 5–19.
- Hempel, C. (1965), *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science* (New York: Free Press).
- Hempel, C. and Oppenheim, P. (1948), 'Studies in the Logic of Explanation', *Philosophy of Science* 15: 135–75.
- Kim, J. (2008), 'Reduction and Reductive Explanation: Is One Possible Without the Other?', in J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation* (New York: Oxford University Press), 93–114.
- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Levy, A. (n.d.), 'Against the Ontic Conception of Explanation'. Manuscript.
- Lewis, D. (1986), 'Causal Explanation', in *Philosophical Papers*, vol. II (New York: Oxford University Press), 214–40.

- Lipton, P. (2009), 'Understanding Without Explanation', in H. W. de Regt, S. Leonelli, and K. Eigner (eds.), *Scientific Understanding: Philosophical Perspectives* (Pittsburgh: University of Pittsburgh Press), 43–63.
- Lombrozo, T. (2011), 'The Instrumental Value of Explanations', *Philosophy Compass* 6: 539–51.
- Norton, J. D. (2012), 'Approximation and Idealization: Why the Difference Matters', *Philosophy of Science* 79: 207–32.
- Pincock, C. (2012), *Mathematics and Scientific Representation* (New York: Oxford University Press).
- Potochnik, A. (2010a), 'Explanatory Independence and Epistemic Interdependence: A Case Study of the Optimality Approach', *British Journal for the Philosophy of Science* 61: 213–33.
- Potochnik, A. (2010b), 'Levels of Explanation Reconceived', *Philosophy of Science* 77: 59–72.
- Potochnik, A. (2015a), 'Causal Patterns and Adequate Explanations', *Philosophical Studies* 172: 1163–82.
- Potochnik, A. (2015b), 'The Diverse Aims of Science', *Studies in History and Philosophy of Science* 53: 71–80.
- Potochnik, A. (2016), 'Scientific Explanation: Putting Communication First', *Philosophy of Science* 83: 721–32.
- Potochnik, A. (2017), *Idealization and the Aims of Science* (Chicago: University of Chicago Press).
- Putnam, H. (1975), 'Philosophy and our Mental Life', in *Philosophical Papers*, vol. II: *Mind, Language and Reality* (Cambridge: Cambridge University Press), 291–303.
- Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World* (Princeton: Princeton University Press).
- Salmon, W. (1989), *Four Decades of Scientific Explanation* (Minneapolis: University of Minnesota Press).
- Strevens, M. (2008), *Depth: An Account of Scientific Explanation* (Cambridge, MA: Harvard University Press).
- Strevens, M. (2013), 'No Understanding Without Explanation', *Studies in History and Philosophy of Science* 44: 510–15.
- van Fraassen, B. C. (1980), *The Scientific Image* (Oxford: Clarendon Press).
- Weslake, B. (2010), 'Explanatory Depth', *Philosophy of Science* 77: 273–94.
- Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation* (New York: Oxford University Press).

4

Extending the Counterfactual Theory of Explanation

Alexander Reutlinger

1. Introduction

The goal of this chapter is to precisely articulate and to extend the counterfactual theory of explanation (CTE). The CTE is a monist account of explanation. I take *monism* to be the view that there is one single philosophical account capturing both causal and non-causal explanations. According to the CTE, both causal and non-causal explanations are explanatory by virtue of revealing counterfactual dependencies between the explanandum and the explanans. I will argue that the CTE is supported by five paradigmatic examples of non-causal explanations in the sciences.

In defending the CTE, I rely on and elaborate recent work of others (see section 2). I also draw on recent work of my own: I apply my version of the CTE (Reutlinger 2016, 2017a) and my Russellian strategy for distinguishing between causal and non-causal explanations (Farr and Reutlinger 2013; Reutlinger 2014) to new examples of non-causal explanations.

As a monist account, the CTE provides *one* philosophical account of *two* types of explanations, around which the recent literature on explanations revolves: causal explanations and non-causal explanations. Examples of causal explanations are familiar instances of causal explanations in the natural and social sciences, including detailed mechanistic explanations (Andersen 2014) and higher-level causal explanations (Cartwright 1989; Woodward 2003; Strevens 2008). Compelling examples of non-causal explanations include different kinds of ‘purely’ or ‘distinctively’ mathematical explanations of contingent phenomena such as graph-theoretic (Pincock 2012, 2015; Lange 2013a), topological (Huneman 2010; Lange 2013a), geometric (Lange 2013a), and statistical explanations (Lipton 2004; Lange 2013b). Other kinds of non-causal explanations are explanations based on symmetry principles and conservation laws (Lange 2011), kinematic principles (Saatsi 2016), renormalization group theory (Batterman 2000; Reutlinger 2014, 2016; Saatsi and Reutlinger forthcoming), dimensional

analysis (Lange 2009; Pexton 2014), functional laws of association or coexistence (Kistler 2013), and inter-theoretic relations (Batterman 2002; Weatherall 2011).¹

The plan of the chapter is as follows: in section 2, I will present three theoretical options to react to examples of non-causal explanations (causal reductionism, pluralism, and monism). In section 3, I introduce and motivate the CTE—a particular kind of monism. In section 4, I argue that the CTE can be successfully applied to five paradigmatic examples of non-causal explanations: Hempel's pendulum explanation (section 4.1), Fermat's explanation (section 4.2), Euler's explanation (section 4.3), renormalization group explanations of universality (section 4.4), and Kahneman and Tversky's explanation (section 4.5). In section 5, I propose the Russellian strategy for distinguishing causal and non-causal explanations. I defend the claim that, if one adopts the Russellian strategy, all five examples of non-causal explanations (presented in section 4) should—in accord with our intuitions—be classified as non-causal. Section 6 provides a conclusion.

Let me add two qualifications: *first*, I will restrict the application of the CTE to examples from the empirical sciences. I will bracket a discussion of non-causal explanations in pure mathematics and philosophy (but see Reutlinger 2017a for an application of the CTE to a class of grounding explanations in metaphysics). *Second*, in this chapter, I will focus on a positive and constructive exposition of the CTE. I address potential worries regarding the CTE elsewhere (Reutlinger 2016, 2017a).

2. Theoretical Options

In this section, I disentangle three distinct strategies for responding to apparent examples of causal and non-causal explanations: (a) causal reductionism, (b) pluralism, and (c) monism. I will, then, provide a *prima facie* reason for defending a monist account.

(a) Causal reductionism is the view that there are no non-causal explanations, because seemingly non-causal explanations can ultimately be understood as causal explanations. Lewis (1986) and, more recently, Skow (2014) have presented one prominent attempt for spelling out this strategy. Typical causal accounts of explanation (such as Salmon 1984; Cartwright 1989; Woodward 2003; Strevens 2008) require identifying the cause(s) of the explanandum. However, Lewis and Skow have weakened the causal account by requiring only that a causal explanation provide some information about the causal history of the explanandum. Lewis's and Skow's notion of causal information is significantly broader than the notion of identifying causes. For instance, Lewis and Skow hold that one causally explains by merely excluding a possible causal history of the explanandum E, or by stating that E has no cause at all, while other causal

¹ I assume here that causal accounts (such as Salmon 1984; Cartwright 1989; Woodward 2003; Strevens 2008) do not provide a general account of all scientific explanations, as causal accounts do not capture non-causal explanations (for details see Reutlinger 2017a: sect. 1, 2017b: sect. 1; van Fraassen 1980: 123; Achinstein 1983: 230–43; Lipton 2004: 32).

accounts would not classify this sort of information as causally explanatory. Lewis and Skow defend the claim that allegedly non-causal explanations (at least, of events, as Skow remarks) turn out to be causal explanations, if one adopts their weakened account of causal explanation.

(b) Pluralism is, roughly put, the view that causal and non-causal explanations are covered by two (or more) distinct theories of explanation. The core idea of a pluralist response to examples of causal and non-causal explanations is that causal accounts of explanations have to be supplemented with an account (or several accounts) of non-causal explanations.

For adopting pluralism, as I define it here, it is, however, not sufficient to merely acknowledge that there are two or more types of explanation—such as causal and non-causal types of explanation. Monists also accept that there are different types of explanations (discussed below). More precisely, a pluralist holds that (1) there are different types of explanations (for present concerns, causal and non-causal types of explanations) and (2) there is no single theory that captures all causal and non-causal explanations, instead one needs two (or more) distinct theories of explanation to adequately capture all causal and non-causal explanations.

Consider two examples of pluralist views.

First, Salmon's claim about the "peaceful coexistence" of the "ontic" causal account and the "epistemic" unification account seems to be an instance of pluralism. Phenomena may have two kinds of explanation: causal "bottom-up" explanations and unificationist "top-down" explanations (Salmon 1989: 183). This is a kind of pluralism because there is no single overarching theory capturing these two types of explanation (Salmon 1989: 184–5).² Instead, Salmon relies on two distinct theories of explanation (a causal account and a unificationist account) to cover certain central cases of causal and non-causal explanations.³

Second, the perhaps most prominent heir of Salmon's pluralist approach in the recent debate on non-causal explanations is Lange's approach (Lange 2011, 2013a, 2016; for an alternative pluralist framework, see Pincock, Chapter 2, this volume). Lange (2013a: 509–10) explicitly refers to Salmon's distinction between "ontic" causal and "modal" theories of scientific explanation. Adopting a modal account, Lange argues that many non-causal explanations operate by showing what constrains the explanandum phenomenon. "Constraining", in this context, amounts to showing why the explanandum had to occur. Lange explicates his modal account in terms of different strengths of necessities: "Distinctively mathematical explanations in science work by appealing to facts [...] that are modally stronger than ordinary causal laws [...]" (Lange 2013a: 491).

² See Reutlinger (2017b) for further details.

³ As a pluralist, Salmon is not committed to the claim that these two accounts cover *all* causal and non-causal explanation. This leaves open the possibility that additional theories of explanation are needed for capturing explanations outside of the scope of causal and unificationist accounts.

Lange is a pluralist, because he agrees with Salmon that (1) there are causal and non-causal types of explanations, (2) there is no overarching, more general account of explanation covering all of these explanations, and some explanations fall under the “ontic” causal account, while some (but not necessarily all) non-causal explanations are subsumed under the “modal” account. Lange summarizes his view: “I have argued that the modal conception, properly elaborated, applies at least to distinctively mathematical explanation in science, whereas the ontic conception does not” (Lange 2013a: 509–10).

(c) **Monism** is the view that there is one single philosophical account capturing both causal and non-causal explanations. A monist holds that causal and non-causal explanations share a feature that makes them explanatory. Unlike the causal reductionist, the monist does not deny the existence of non-causal explanations. The monist disagrees with the pluralist, because the former wishes to *replace* causal accounts of explanation with some monist account (for instance, the CTE), while the latter wants to *supplement* causal accounts with a theory of non-causal explanations.

Hempel’s covering-law account is an instructive historical example for illustrating monism (Hempel 1965: 352). Hempel argues that causal and non-causal explanations are explanatory by virtue of having one single feature in common: nomic expectability of the explanandum. In the case of causal explanations, one expects the explanandum to occur on the basis of causal covering laws (laws of succession) and initial conditions; in the non-causal case, one’s expectations are based on non-causal covering laws (laws of coexistence) and initial conditions. However, Hempelian monism is unfortunately not the most attractive option for monists, because his covering-law account suffers from well-known problems (Salmon 1989: 46–50).

Currently, it is an open question as to whether there is a viable monist alternative to Hempelian monism (Lipton 2004: 32). The perhaps most promising and the most elaborate recent attempt to make progress on a monist approach are counterfactual theories of causal and non-causal explanations. Proponents of the counterfactual theory have articulated and explored this approach in application to various examples of non-causal explanations (Frisch 1998; Bokulich 2008; Kistler 2013; Saatsi and Pexton 2013; Pexton 2014; Pincock 2015; Rice 2015; Reutlinger 2016, 2017a; Saatsi 2016; French and Saatsi, Chapter 9, this volume; Woodward, Chapter 6, this volume).⁴

I have presented three theoretical options to react to the existence of causal and non-causal explanations. Here and elsewhere, I articulate and defend the CTE as a monist approach. But why should one opt for monism rather than for pluralism or causal reductionism? What is so attractive about monism? The answer is straightforward: *prima facie*, monism is superior to the alternative theoretical options for two reasons. Firstly, there are compelling examples of what seem to be non-causal explanations in the sciences (section 1). Monism is superior to causal reductionism because the former

⁴ Mach (1872: 35–7) anticipates current counterfactual accounts.

allows for the existence of non-causal explanations, while the latter does not adequately capture these examples of scientific explanations. Secondly, *ceteris paribus*, philosophers prefer more general philosophical theories to less general theories. Given this preference, monism is superior to pluralism because the former provides *one* general theory of causal and non-causal explanations in science, while pluralist construals consist of two or more theories. For these reasons, I take it that monism is an attractive view deserving further exploration.

3. The Counterfactual Theory of Explanation

Current counterfactual theories take Woodward's counterfactual account of causal explanations as their starting point:

An explanation ought to be such that it enables us to see what sort of difference it would have made for the explanandum if the factors cited in the explanans had been different in various possible ways. (Woodward 2003: 11)

Explanation is a matter of exhibiting systematic patterns of counterfactual dependence. (Woodward 2003: 191)

Woodward's version of the counterfactual theories of explanation and its underlying interventionist theory of causation is originally intended to capture causal explanations (Woodward 2003: 203). However, the core idea of the counterfactual theory—that is, analyzing explanatory relevance in terms of *counterfactual dependence*—is not necessarily tied to a causal interpretation. Woodward suggests this line of argument, although without pursuing this intriguing idea any further (but see Woodward, Chapter 6, this volume):⁵

[T]he common element in many forms of explanation, both causal and non-causal, is that they must answer what-if-things-had-been-different questions. (Woodward 2003: 221)

To answer what-if-things-had-been-different questions is to reveal how the explanandum counterfactually depends on possible changes in the initial conditions (that are part of the explanans). The monist proposal of the CTE is that causal and non-causal explanations are explanatory by virtue of exhibiting how the explanandum counterfactually depends on the explanans.

I will now provide a more precise characterization of the CTE⁶ in terms of the following necessary conditions:

1. **STRUCTURE CONDITION:** Explanations have a two-part structure consisting of a statement E about the occurrence of the (type or token of the) explanandum

⁵ See Lipton (2004: 32) regarding a similar approach.

⁶ I follow Woodward's (2003: 203) and Woodward and Hitchcock's (2003: 6, 18) exposition of the CTE, building on Reutlinger (2016, 2017a).

phenomenon; and an explanans including nomic⁷ generalizations $G_1, \dots, G_m$, statements about initial (or boundary) conditions $IC_1, \dots, IC_n$, and, typically, further auxiliary assumptions $A_1, \dots, A_o$ (such as Nagelian bridge laws, symmetry assumptions, limit theorems, and other modeling assumptions).

2. VERIDICALITY CONDITION: $G_1, \dots, G_m$, $IC_1, \dots, IC_n$, $A_1, \dots, A_o$, and E are (approximately) true.
3. INFERENCE CONDITION: $G_1, \dots, G_m$ and $IC_1, \dots, IC_n$ allow us to deductively infer E , or to infer a conditional probability $P(E|IC_1, \dots, IC_n)$. This conditional probability need not be high, in contrast to Hempel's covering-law account; it is merely required that $P(E|IC_1, \dots, IC_n) > P(E)$.
4. DEPENDENCY CONDITION: $G_1, \dots, G_m$ support at least one counterfactual of the form: if the initial conditions $IC_1, \dots, IC_n$ had been different than they actually are (in at least one specific way deemed possible in the light of the nomic generalizations), then E , or the conditional probability of E , would have been different as well.⁸

In sum, the CTE is a monist view because causal and non-causal explanations are explanatory in virtue of satisfying these conditions.

Let me add three qualifications:

First qualification. It is reasonable to require a fifth necessary condition to be in place, which is, jointly with the other four, sufficient: that is, the MINIMALITY CONDITION, according to which no proper subset of the set of explanans statements $\{G_1, \dots, G_m, IC_1, \dots, IC_n, A_1, \dots, A_o\}$ satisfies all of conditions 1–4 of the CTE. Due to space constraints, I will not discuss this condition explicitly when applying the CTE (section 4). I will simply assume that it is satisfied for examples of scientific explanations. The main purpose of the MINIMALITY CONDITION is to guard against including irrelevant factors into the explanans, which constitutes a familiar problem for Hempel's monism (Salmon 1989: 50). However, as already emphasized, I will not discuss potential objections to the CTE here (Reutlinger 2016, 2017a).

Second qualification. Although the VERIDICALITY CONDITION is met in the case of *some* scientific explanations, one might worry that the VERIDICALITY CONDITION does not hold for all scientific explanation, because (a) many scientific explanations

⁷ I require that the generalization be nomic mainly because I assume that only nomic generalizations support counterfactuals (see the dependency condition below). I use a broad notion of laws that includes non-strict ceteris paribus laws, such as Woodward's (2003) own invariance account. However, my aim here is not to defend a particular view of laws. The CTE is neutral with respect to alternative theories of lawhood, which is a strength of the CTE.

⁸ I speak of nomic generalizations "supporting" or "underwriting" counterfactuals. These expressions serve as a proxy for a precise semantics for (causal and non-causal) counterfactuals. Prima facie, none of the major approaches to the meaning of counterfactuals is ruled out for the CTE when applied to non-causal explanations, such as Goodmanian approaches, possible worlds semantics, and suppositionalist accounts (Bennett 2003). It is a task for future research to explore these alternative semantic approaches within the CTE framework.

involve idealized (auxiliary) assumptions, and (b) how-possibly explanations play an important epistemic role in the sciences. Both idealized and how-possibly explanations do not meet the veridicality condition, or that is the worry. Regarding idealized explanations, it is, however, often possible to (re)interpret the idealizations in a way that is compatible with the veridicality condition by adopting, for instance, dispositionalist and minimalist accounts of idealizations (Cartwright 1989; Hüttemann 2004; Strevens 2008). Regarding how-possibly explanations, I ultimately agree that the veridicality condition has to be rejected, if the CTE is supposed to be an account of both how-possibly and how-actually explanations. However, many prominent accounts of explanations (including Woodward's CTE) are at least implicitly presented as accounts of how-actually explanations. In this vein, I also introduce the CTE as an account of how-actually explanations. This account can, of course, be weakened in the case of how-possibly explanations (Reutlinger et al. 2017).

Third qualification. Woodward interprets the counterfactuals figuring in the DEPENDENCY CONDITION in terms of a specific counterfactual theory of causation, his interventionist theory. For Woodward, causal counterfactuals just are interventionist counterfactuals. But the dependency condition can, to a certain extent, be disentangled from (I) an interventionist account of causation (in the context of causal explanations), and from (II) a causal interpretation (in the context of non-causal explanations). Let me explain the claims (I) and (II) in more detail.

Consider claim (I) first. Although Woodward interprets the dependency condition causally in terms of interventionist counterfactuals, a proponent of the CTE is not committed to an interventionist account of causation. Other broadly counterfactual accounts of causation are also compatible with the CTE (including, for instance, von Wright 1971; Lewis 1973; Menzies and Price 1993; Reutlinger 2013: chapter 8). In this chapter, it is not my goal to argue for any particular counterfactual account of causation. As a consequence, I will not commit myself to the claim that—in the context of *causal* explanation—the counterfactuals mentioned in the dependency condition of the CTE have to be understood as interventionist counterfactuals.

Let me now turn to claim (II) that the dependency condition can be disentangled from a *causal* interpretation. Woodward himself voices a *prima facie* convincing reason for not requiring that *all* explanatory counterfactuals have the form of interventionist counterfactuals:

When a theory or derivation answers a what-if-things-had-been-different question but we cannot interpret this as an answer to a question about what would happen under an intervention, we may have a non-causal explanation of some sort. (Woodward 2003: 221)

As I understand this quote, Woodward draws a distinction between causal (for him, interventionist) counterfactuals and non-causal (for him, non-interventionist) counterfactuals, both of which can be exploited for explanatory purposes. That is, while causal explanations rely on interventionist counterfactuals, there are also non-causal explanations making use of non-interventionist counterfactuals.

The idea of introducing a distinction between causal and non-causal counterfactuals is not necessarily restricted to Woodward's interventionist version of the CTE. It applies to the CTE more generally. Unlike Woodward, proponents of non-interventionist, broadly counterfactual theories of causation draw the line separating causal and non-causal counterfactuals differently, i.e., not necessarily in terms of interventions. In section 4, I suggest to distinguish between causal and non-causal counterfactuals on the basis of Russellian criteria.

4. Applying the Counterfactual Theory of Explanation

I will now argue for the claim that the CTE applies to five paradigmatic examples of non-causal explanations. I will start with two instructive examples of non-causal explanations that Hempel introduced into the debate: Hempel's pendulum explanation (section 4.1) and Fermat's explanation (section 4.2). Then, I will apply the CTE to examples from the more recent literature: Euler's explanation (section 4.3), renormalization group explanations of universality (section 4.4), and, finally, to Kahneman and Tversky's explanation (section 4.5).⁹

4.1 Hempel's pendulum explanation

Let us take Hempel's (1965: 352) instructive example of a non-causal explanation as a starting point:

[...] D-N explanations are not always causal. For example, the fact that a given simple pendulum takes two seconds to complete one full swing might be explained by pointing out that its length is 100 centimeters, and the period t (in seconds), of any simple pendulum is connected with its length l (in centimeters) by the law [of the simple pendulum, A.R.]. (Hempel 1965: 352)

Call this 'Hempel's pendulum explanation'. Hempel considers this explanation to be representative of a class of scientific explanations that rest on *laws of coexistence* (for further examples of "association laws", Kistler 2013: 68–71). Hempel claims that his covering-law model applies to this example, since the occurrence of the explanandum was to be expected on the basis of the law of the simple pendulum and the initial conditions.

However, my present concern is whether the CTE captures Hempel's pendulum explanation. I argue that the CTE is applicable to Hempel's pendulum explanation.

First, Hempel's pendulum explanation satisfies the STRUCTURE CONDITION required by the CTE. The *explanandum* statement refers to the phenomenon that some particular simple pendulum actually takes two seconds to complete one full swing. The *explanans* of Hempel's pendulum explanation consists of (a) the law of the simple pendulum, (b) a statement about the initial conditions including that the pendulum has been set into motion and that the length of the pendulum is 100 centimeters, and (c) further

⁹ In sections 4.3 and 4.4 I use material from Reutlinger (2016).

auxiliary background assumptions such as that there is no air resistance (although Hempel does not mention them explicitly).

Second, the VERIDICALITY CONDITION is met, since the statements about initial conditions and the law statement are approximately true.¹⁰

Third, the INFERENCE CONDITION is satisfied because the explanans statements logically entail the explanandum statement.

Finally, the DEPENDENCY CONDITION is satisfied, as the law of the simple pendulum underwrites counterfactuals such as: if the length of the pendulum had been different from 100 centimeters, then its period would have been different (i.e., the simple pendulum would have taken more or less than two seconds to complete one full swing).

Thus, the CTE is applicable to Hempel's pendulum explanation.

4.2 *Fermat's explanation*

Hempel's second example of a non-causal explanation is a typical instance of explanations drawing on *variational principles* (such as Fermat's principle of least time, the principle of least action, Gauss's principle, and Hertz's principle). Hempel's example is interesting particularly because the explanatory role of variational principles has been, by and large, neglected in the recent debate on (non-causal) explanations (van Fraassen 1989: 234 and Lange 2016: 68 are welcome exceptions).¹¹

Hempel describes the following example of an explanation based on Fermat's principle of least time:

Consider, for example, a beam of light that travels from a point A in one optical medium to a point B in another, which borders upon the first along a plane. Then, according to Fermat's principle of least time, the beam will follow a path that makes the traveling time from A to B a minimum as compared with alternative paths available. Which path this is will depend on the refractive indices of the two media; we will assume that these are given. Suppose now that the path from A to B determined by Fermat's principle passes through an intermediate point C.

Hempel argues that one now has the means to explain why the beam of light passed through point C:

[T]his fact may be said D-N explainable by means of Fermat's law in conjunction with the relevant data concerning the optical media and the information that the light traveled from A to B. (Hempel 1965: 353)

Let us call this explanation 'Fermat's explanation.' Hempel holds that the covering-law account captures Fermat's explanation, because the beam passing through point C was to be expected on the basis of Fermat's principle and the initial conditions.

Does the CTE apply to Fermat's explanation? I will argue that it does.

¹⁰ I will simply assume that there is an interpretation of the idealized assumption that there is no air resistance satisfying the veridicality condition. I will not discuss the issue of idealizations in this chapter.

¹¹ See also Yourgrau and Mandelstam (1968).

First, Fermat's explanation satisfies the STRUCTURE CONDITION. The *explanandum* is a statement about the fact that the beam of light passed through a point C—an intermediate point between points A and B. The *explanans* consists of Fermat's principle of least time (a nomic generalization), the statement about initial conditions that a beam of light traveled from a point A in one optical medium to a point B in another medium, and further assumptions about the optical media (such as “the refractive indices of the two media”).

Second, the explanation meets the VERIDICALITY CONDITION, since the explanans consisting of Fermat's principle of least time, the statement that a beam of light that traveled from a point A (at t_1) in one optical medium to a point B (at t_3) in another medium (plus further assumptions about the optical media) is approximately true. Furthermore, it is true that the beam of light passes through point C at t_2 (the explanandum statement).

Third, it satisfies the INFERENCE CONDITION, as the explanans entails the explanandum statement.

Fourth, the DEPENDENCY CONDITION is met because Fermat's principle allows us to evaluate the following counterfactuals as true: (i) ‘if the beam had traveled from point A t_1 to point B* at t_3 (in contrast to point B), the beam would have gone through point C* at t_2 (in contrast to point C)’ and (ii) ‘if the beam had traveled from point A* at t_1 (in contrast to point A) to point B at t_3 , the beam would have gone through point C** at t_2 (in contrast to point C)’.

Therefore, the CTE applies to Fermat's explanation.

4.3 Euler's explanation

Euler's explanation is an intuitively simple and powerful non-causal “graph-theoretical” explanation (van Fraassen 1989: 236–9; Pincock 2012: 51–3; Lange 2013a: 489; Reutlinger 2016: 730–40; Jansson and Saatsi forthcoming). I use Euler's explanation as a stand-in for graph-theoretical and network-based explanations (see Huneman 2010).

In 1736, Königsberg had four parts of town and seven bridges connecting these parts. Interestingly, no one, at that time, ever succeeded in the attempt to cross all of the bridges exactly once. This surprising fact calls for an explanation. The mathematician Leonhard Euler provided an explanation. Euler's explanation starts with representing relevant aspects of Königsberg's geography with a graph. A simplified geographical map of Königsberg in 1736 represents only the four parts of town (the two islands A and B, and the two riverbanks C and D) and the seven bridges (part A is connected to five bridges, parts B, C, and D are each connected to three bridges). This simplified geography of Königsberg can also be represented by a graph, in which the nodes represent the parts of town A–D and the edges represent the bridges.

Relying on this graph-theoretical representation, Euler defines an Euler path as a path through a graph G that includes each edge in G exactly once. Euler uses the notion of an Euler path to reformulate the explanandum in terms of the question: why has

everyone failed to traverse Königsberg on an Euler path? His answer to this why-question has two components.

First, Euler's theorem according to which there is an Euler path through a graph G iff G is an Eulerian graph. Euler proved that a graph G is Eulerian iff (i) all the nodes in G are connected to an even number of edges, or (ii) exactly two nodes in G (one of which we take as our starting point) are connected to an odd number of edges.

Second, the actual bridges and parts of Königsberg are *not* isomorphic to an Eulerian graph, because conditions (i) and (ii) in the definition of an Eulerian graph are not satisfied: no part of town (corresponding to the nodes) is connected to an even number of bridges (corresponding to the edges), violating condition (i); and more than two parts of town (corresponding to the nodes) are connected to an odd number of bridges (corresponding to the edges), violating condition (ii). Königsberg could have been isomorphic to an Eulerian graph in 1736, but as a matter of *contingent fact* it was not.

Therefore, Euler concludes from the first and the second component that there is no Euler path through the actual Königsberg. This explains why nobody ever succeeded in crossing all of the bridges of Königsberg exactly once.

Does the CTE capture Euler's explanation? All four *conditions* that the CTE imposes on the explanans and the explanandum are satisfied:

First, Euler's explanation is in accord with the STRUCTURE CONDITION. The *explanandum* phenomenon is the fact that everyone has failed to cross the city on an Euler path. The *explanans* consists of Euler's theorem (a mathematical and intuitively non-causal generalization concerning graphs) and a statement about the contingent initial conditions that all parts are actually connected to an odd number of bridges.

Second, the VERIDICALITY CONDITION holds because (a) Euler's theorem, (b) the statement about the contingent fact that each part of Königsberg is actually connected to an odd number of bridges, and (c) the explanandum statement are all true.

Third, the INFERENCE CONDITION is met, since Euler's theorem together with the statement about the contingent initial conditions entail the explanandum statement.

Fourth, the DEPENDENCY CONDITION is satisfied, because Euler's theorem supports counterfactuals such as: (i) 'if all parts of Königsberg had been connected to an *even* number of bridges, then people would *not* have failed to cross all of the bridges exactly once', and (ii) 'if exactly two parts of town were connected to an *odd* number of bridges, then people would *not* have failed to cross all of the bridges exactly once'.¹²

Therefore, I conclude that the CTE applies to Euler's explanation.

4.4 Renormalization group explanations

Renormalization group (RG) explanations are intended to explain why microscopically different physical systems display the same macro-behavior when undergoing phase-transitions. For instance, near the critical temperature, the phenomenology of

¹² I am assuming that, in these counterfactual situation(s), the inhabitants of Königsberg are intelligent and try repeatedly to walk over all of bridges exactly once.

transitions of a fluid from a liquid to a vaporous phase, or of a metal from a magnetic to a demagnetized phase is (in some respects) the same, although liquids and metals are significantly different on the micro-level. This ‘sameness’ or—to use a more technical term—‘universality’ of the macro-behavior is characterized by a critical exponent that takes the same value for microscopically very different systems (Batterman 2000: 125–6). How do physicists explain the remarkable fact that there is universal macro-behavior?

It is useful to reconstruct RG explanations as involving three key explanatory elements: (1) Hamiltonians, (2) RG transformations, and (3) the flow of Hamiltonians. There is a fourth element—the laws of statistical mechanics, including dynamical laws and the partition function—which I will leave in the background, for brevity’s sake (Norton 2012: 227; Wilson 1983). The exposition of these elements will be non-technical because the chapter is concerned with a non-technical question (for a more detailed exposition see Fisher 1982, 1998; Wilson 1983; Saatsi and Reutlinger forthcoming).

- (I) *Hamiltonians*: The Hamiltonian is a function characterizing, among other things, the energy of the interactions between the components of the system. One characteristic of a physical system undergoing a (continuous) phase transition is that the correlation length diverges and becomes infinite. That is, the state of every component becomes correlated not only with the states of its nearby components but also with the states of distant components. The correlation length diverges, although each component interacts merely locally with its nearby neighbors (Batterman 2000: 126, 137–8). Adopting Batterman’s (2000) terminology, I call this complicated Hamiltonian of a system undergoing a phase transition the “original” Hamiltonian.
- (II) *Renormalization group transformations*: Keeping track of the correlations and interactions between the components of a system undergoing a phase transition is—given the large number of components and the diverging correlation length—practically impossible. So-called renormalization group transformations (henceforth, RG transformations) deal with this intractability by redefining the characteristic length, at which the interactions among the components of the system at issue are described. Repeatedly applying RG transformations amounts to a re-description of the system, say fluid F, on larger and larger length scales while preserving the mathematical form of the original Hamiltonian. The transformed Hamiltonian describes a system (and the interactions between its components) with less degrees of freedom than the original Hamiltonian. In sum, the RG transformation is a mathematically sophisticated coarse-graining procedure eliminating micro-details that are irrelevant for the explanation of universality.
- (III) *The flow of Hamiltonians*: Suppose we start with the original Hamiltonian H of a fluid F undergoing a phase transition. Then, one repeatedly applies the RG transformation and obtains other more ‘coarse-grained’ Hamiltonians.

Interestingly, these different Hamiltonians “flow” into a fixed point in the space of possible Hamiltonians, which describes a specific behavior characterized by a critical exponent (Batterman 2000: 143). Now suppose there is another fluid F^* and its behavior (during phase transition) is described by the initial Hamiltonian H^* . Repeatedly applying the RG transformation to H^* generates other, more ‘coarse-grained’ Hamiltonians. If the Hamiltonians representing fluid F^* and fluid F turn out to “flow” to the same fixed point, then their behavior, when undergoing phase transition, is characterized by the same critical exponent (Fisher 1982: 85; Batterman 2000: 143).

The three elements of an RG explanation allow us to determine whether systems with different original Hamiltonians belong to the same “universality class” and are characterized by the same critical exponent (Fisher 1982: 87). Two systems belong to the same universality class because reiterating RG transformations reveals that both systems “flow” to the same fixed point.

Now the decisive question is whether the CTE applies to RG explanations.

First, RG explanations exhibit the form required by the STRUCTURE CONDITION. The *explanandum* phenomenon is the occurrence of universal macro-behavior. The *explanans* of an RG explanation consists of the system-specific Hamiltonians describing the energy state of the physical systems in question—and, strictly speaking, the laws of statistical mechanics (the fourth element in the background); RG transformations and the flow of Hamiltonians are central auxiliary assumptions in the explanans.

Second, for the purpose of this chapter, I will take the VERIDICALITY CONDITION to be satisfied, because the explanandum statement (that there is universal behavior) and the explanans can—for present purposes—be considered as being (approximately) true. Due to space limitations, I cannot discuss the role of idealizations (especially, limit theorems) in RG explanations posing a potential threat to the truth of the explanans. However, there are interpretations of the idealizations in question that are consistent with the veridicality condition (see Strevens 2008; Norton 2012; Saatsi and Reutlinger forthcoming).

Third, the INFERENCE CONDITION holds, since the RG explanans entails that many physical systems with different original Hamiltonians display the same macro-behavior.

Fourth, the DEPENDENCY CONDITION is met, because the RG explanans supports *some* counterfactuals of the form: ‘There is a physically possible Hamiltonian H^* such that: if a physical system S had the original Hamiltonian H^* (instead of its actual original Hamiltonian H), then S with original H^* would be in a different universality class than a system with original Hamiltonian H .’

Let me clarify why the dependency condition holds in the light of RG theory. The main accomplishment of RG explanations is to show that many systems with different original Hamiltonians belong to the *same* universality class. However, the dependency condition of the CTE does not require that the explanandum depend on *all* possible changes in the initial conditions. Instead the condition merely requires that

the explanandum counterfactually depend on *some* possible changes in the explanans. Indeed, RG theory also shows that and why some systems with different original Hamiltonians do *not* exhibit the same macro-behavior and in fact belong to *different* universality classes (Wilson 1983). As Batterman (2000: 127) points out, RG explanations reveal that belonging to a particular universality class *depends* on features of the physical system such as the symmetry properties of the order parameter, the spatial dimensionality, and the range of the microscopic interactions. One can express this dependency with the following counterfactuals:

- If a physical system *S* had a different spatial dimensionality than it actually has, then *S* would be in a different universality class than it actually is in.
- If a physical system *S* had a different symmetry of the order parameter than it actually has, then *S* would be in a different universality class than it actually is in.
- If a physical system *S* had a (sufficiently) different range of the microscopic interactions than it actually has, then *S* would be in a different universality class than it actually is in.

Hence, if systems with H^* and H —figuring in the counterfactual above—differ with respect to those features, then the counterfactuals are true, according to RG theory (Reutlinger 2016; Saatsi and Reutlinger forthcoming).

Therefore, I conclude that the CTE successfully captures RG explanations.

4.5 Kahneman and Tversky's explanation

Kahneman and Tversky's explanation is representative of a larger class of non-causal statistical explanations (Lange 2013b, 2016 discusses further examples of non-causal statistical explanations). Following Lipton's exposition, Kahneman and Tversky's explanation is concerned with the following phenomenon:

Flight instructors in the Israeli air force had a policy of strongly praising trainee pilots after an unusually good performance and strongly criticizing them after an unusually weak performance. What they found is that trainees tended to improve after a poor performance and criticism; but they actually tended to do worse after good performance and praise. (Lipton 2004: 32)

What explains this phenomenon? After briefly considering a causal explanation, Kahneman and Tversky propose the following non-causal answer:

Perhaps it is that criticism is much more effective than praise. That would be a causal explanation. But this pattern is also what one should expect if neither praise nor criticism had any effect. It may just be regression to the mean: extreme performances tend to be followed by less extreme performances. If this is what is going on, we can have a lovely explanation of the observed pattern by appeal to chance (or the absence of causal influence) rather than any cause. (Lipton 2004: 32)

Does the CTE capture Kahneman and Tversky's explanation?

First, the explanation satisfies the STRUCTURE CONDITION. I take the *explanandum* phenomena to be short sequences of particular events such as (1) that pilot P_1 's first

flight performance was extremely good, P_1 was strongly praised for it, but her second flight performance was worse than the first, and (2) that pilot P_2 's first flight performance was extremely poor, P_2 was strongly criticized for it, and her second flight performance was better than the first. The *explanans* consists of a statistical generalization stating that extreme performances tend to be followed by less extreme performances. More generally put, the statistical generalization states that (a measurement of) extreme values of a variable tend to be followed by (a measurement of) less extreme values of that variable, i.e. values that are closer to the mean.¹³ The initial conditions in this example express the outcome of the first flight performance of a given pilot (for instance, pilot P_1 's first flight performance was extremely good) and whether the pilot was strongly praised or criticized afterwards.

Second, the VERIDICALITY CONDITION is met because the explanandum statement, the statistical generalization and the statements about actual performances (and praise/criticism) are approximately true.

Third, the explanation satisfies the INFERENCE CONDITION since the explanans implies a conditional probability for the explanandum phenomenon (although the probabilities are vague in this example, as the expression "tend to" indicates). For instance, the statistical generalization and the information that pilot P_1 's first flight performance was extremely good (and P_1 was strongly praised for it) allow us to infer that it is highly probable that P_1 's second flight performance will be worse than the first.

Fourth, the DEPENDENCY CONDITION holds, because the statistical generalization supports the following two counterfactuals: (i) regarding the first explanandum, 'if P_1 's first performance had been extremely poor (as it actually was not), then the probability would have been high that P_1 does better in the second performance than in the first performance', and (ii) regarding the second explanandum, 'if P_2 's first performance had been extremely good (as it actually was not), then the probability would be high that P_2 does worse in the second performance than in the first performance'.

Thus, the CTE captures Kahneman and Tversky's explanation.

5. Distinguishing Non-Causal and Causal Explanations: The Russellian Strategy

So far I have taken it for granted that the examples presented and discussed in section 3 are indeed non-causal explanations. But what makes them non-causal? How does one distinguish between causal and non-causal explanations? I propose a 'Russellian' strategy for distinguishing between causal and non-causal explanations within the CTE framework. The Russellian strategy involves two steps.

¹³ One may ask for a (mathematical) explanation of this statistical principle but this is not the topic here (see Lange 2013b).

First step. Following Bertrand Russell (1912/13) and present-day Neo-Russellians (Field 2003; Ladyman and Ross 2007; Norton 2007; Farr and Reutlinger 2013; Reutlinger 2013, 2014; also Frisch 2014), I use the following criteria to characterize causal relations:

- asymmetry (that is, if A causes B, then B does not cause A),
- time asymmetry (that is, causes occur earlier than their effects),¹⁴
- distinctness of the causal relata (that is, cause and effect do not stand in a part-whole, supervenience, grounding, determinable-determinate, or any other metaphysical dependence relation),
- metaphysical contingency (that is, causal relations obtain with metaphysical contingency).

I will refer to these criteria as ‘Russellian criteria.’ The mentioned Russellian criteria are taken to be necessary (but not sufficient), or at least typical, conditions for causation. Adopting a broadly counterfactual theory of causation, I assume that counterfactual dependencies deserve a causal interpretation only if (or, more cautiously, to the extent to which) the dependencies have all of the Russellian features.¹⁵

Second step. We can now use the Russellian criteria to distinguish between causal and non-causal explanations within the framework of the CTE. The key idea is that not all explanatory counterfactuals are alike. Causal explanations are explanatory by virtue of exhibiting causal counterfactual dependencies; non-causal explanations are explanatory by virtue of exhibiting non-causal counterfactual dependencies. Taking into account the Russellian criteria, causal explanations reveal causal counterfactual dependencies if the dependency relations satisfy *all* of the Russellian criteria. Non-causal explanations exhibit non-causal counterfactual dependencies, if the dependency relations do not satisfy all of the Russellian criteria.

I will now apply the Russellian strategy to argue that all of the examples discussed in section 4 are instances of non-causal explanations. Finally, I will conclude the section with a general remark on the asymmetry of non-causal explanations.

(a) Hempel’s pendulum explanation. Hempel argues that the explanation is non-causal because the covering law (the law of the simple pendulum) is a non-causal law of coexistence:

This law [i.e., the law of the pendulum] expresses a mathematical relationship between the length and the period (which is a quantitative dispositional characteristic) of the pendulum *at one and the same time*. (Hempel 1965: 352; emphasis added)

[L]aws of this kind, of which the laws of Boyle and of Charles as well as Ohm’s law are other examples, are sometimes called *laws of coexistence*, in contradistinction to *laws of succession*,

¹⁴ I will not address the possibility of backwards causation in the domain of theories in fundamental physics. I merely assume that time asymmetry is a typical feature of causation in non-fundamental physics and in the special sciences (Albert 2000; Loewer 2007; Reutlinger 2013; Frisch 2014).

¹⁵ Advocates of broadly counterfactual accounts of causation tend to accept the Russellian criteria (Lewis 1973, 1979; Albert 2000; Elga 2001; Woodward 2003, 2007; Loewer 2007).

which concern the temporal change of a system. These latter include, for example, Galileo's law and the laws for the change of state in systems covered by a deterministic theory. Causal explanation by reference to the antecedent events clearly presupposes laws of succession; in the case of the pendulum, where only a law of coexistence is invoked, one surely would not say that the pendulum's having a period of two seconds was caused by the fact that it had a length of 100 centimeters. (Hempel 1965: 352)

Hempel regards explanations based on laws of coexistence as non-causal because—using my terminology—such explanations lack (at least) one Russellian criterion of causality: they are not time-asymmetric, since laws of coexistence relate physical quantities “at one and the same time” (Hempel 1965: 352).

Hempel's claim can easily be reformulated in the framework of the CTE. Hempel's pendulum explanation rests on counterfactual dependencies of the period of the pendulum on its length. Supposing that Hempel correctly asserts that the law of the pendulum is a law of coexistence, the relevant counterfactual dependencies are not time-asymmetric, since the dependence holds between physical states (length and period) at one and the same time. Hence, the counterfactual dependencies of Hempel's pendulum explanation lack at least one Russellian criterion, time asymmetry. Thus, the explanation is non-causal (see Kistler 2013 for further examples of counterfactual dependencies based on laws of coexistence).

(b) Fermat's explanation. Hempel suggests that the character of Fermat's explanation is non-causal due to a lack of time asymmetry. But the violation of time asymmetry in that case differs from the lack of time asymmetry in the case of the pendulum explanation (Hempel 1965: 353). Explaining why the beam of light passes through point C at t_2 (on the basis on Fermat's principle) refers to an earlier event (the beam passing through point A at t_1) and also to a *later* event (the beam passing through point B at t_3). Hempel argues that explanatory reference to an event occurring later than the explanandum event violates time asymmetry.

Agreeing with Hempel's diagnosis, one can reformulate this point in terms of the CTE. Recall one relevant counterfactual in the context of Fermat's explanation: ‘if the beam had traveled from point A at t_1 to point B* at t_3 (in contrast to point B at t_3), it wouldn't have gone through point C at t_2 ’. This counterfactual is not time-asymmetric, because the antecedent refers to an event occurring earlier and also to another event occurring later than the explanandum event. Thus, Fermat's explanation is non-causal because it does not instantiate at least one of the Russellian criteria.

(c) Euler's explanation. The explanation is non-causal because it lacks several Russellian criteria. First, the relevant counterfactual dependencies (between numbers of bridges per part of town and the ability to cross the bridges) are not time-asymmetric. In the context of Euler's explanation, the fact that Königsberg instantiates a certain graph-theoretical structure does not occur earlier than the failed attempts to cross the bridges—at least not in any sense relevant for the explanation. It is rather a presupposition

of Euler's explanation that Königsberg does not change its structure during the entire course of attempted bridge-crossings. Second, the explanans facts (including that Königsberg actually instantiates a certain kind of graph and that people actually attempted to cross the bridges) and the explanandum fact (that is, people failing to cross each bridge exactly once) are—unlike facts about causes and effects—not distinct facts. Distinct facts are defined as facts that do not stand in a part–whole, supervenience, grounding, determinable–determinate, or any other metaphysical dependence relation. In the case of Euler's explanation, the explanans facts and the explanandum fact are not distinct, because the explanandum fact that people fail to cross each bridge exactly once supervenes on (or metaphysically depends on, or is grounded in) the explanans fact that Königsberg instantiates a particular kind of graph (and the fact that people actually attempted to cross the bridges).¹⁶ Third, Euler's explanation lacks metaphysical contingency. It is metaphysically, or mathematically, impossible (and not merely physically impossible) to cross the bridges as planned, if Königsberg instantiates a non-Eulerian graph (see Lange 2013a; Reutlinger 2014; Andersen forthcoming). In sum, Euler's explanation lacks at least three Russellian criteria. Hence, it is a non-causal explanation.

(d) RG explanations of universality. The RG explanation of universality is non-causal because it lacks, at least, two Russellian criteria (see Reutlinger 2014).

First, the relevant counterfactuals are not time-asymmetric, because RG transformations relate Hamiltonians, but the original Hamiltonian H does not occur before (or after) any of the transformed Hamiltonian H^* .

Second, the RG counterfactuals do not relate distinct events. The RG transformations relate Hamiltonians, but these Hamiltonians do not represent distinct states of physical systems. Instead, H and H^* represent the different degrees of coarse-grained representations of the same physical system. Furthermore, having a particular critical exponent, or belonging to a particular universality class (as the consequent of a relevant RG counterfactual states) is a macroscopic feature supervening or metaphysically depending on (at least partially) microscopic facts (such as that a physical system has a specific original Hamiltonian that is subject to RG transformations, and on features such as symmetry properties of the order parameter and the spatial dimensionality, and the range of the microscopic interactions).

(e) Kahneman and Tversky's explanation. Following Lange, I take it that the statistical generalization ("regression to the mean") is a mathematical truth, a "statistical fact of life" (Lange 2013b: 173). If that is correct, then the explanation is non-causal because its main generalization and the counterfactual dependencies this generalization underwrites lack metaphysical contingency, one of the Russellian criteria. Moreover,

¹⁶ One might worry that the explanatory facts are identical with the fact to be explained, if one does not require distinctness. However, asserting that two facts are not distinct does not imply that they are identical (for instance, two facts might not be distinct because one fact supervenes on, or is grounded, in the other).

although the relevant counterfactuals refer to *earlier* extreme performances in the antecedent and to *later* non-extreme performances in the consequent, this time-asymmetric order does not seem to be essential for the explanation. One could also use “regression to the mean” to explain why the later flight performance was extreme and the earlier performance was not. Hence, it is not obvious that Kahneman and Tversky’s explanation and the relevant counterfactual dependencies are time-asymmetric. In sum, Kahneman and Tversky’s explanation qualifies as being non-causal.

(f) Are non-causal explanations asymmetric? Causal explanations may well all be asymmetric. But is this also true of all non-causal explanations? I claim that some, but not necessarily all, non-causal explanations lack the Russellian criterion of *asymmetry* in that the counterfactual dependence in question is symmetric.¹⁷

Let me briefly motivate why I think that some non-causal explanations are not asymmetric. Let us define the notion of an explanation being asymmetry. An explanation is asymmetric iff (1) the initial conditions and the nomic generalizations explain the explanandum, and (2) it is not the case that explanandum and the nomic generalizations explain the initial conditions. If one relies on the CTE, an explanation is asymmetric (I focus on the DEPENDENCY CONDITION for ease of presentation) only if (1) the explanandum counterfactually depends on the initial conditions (given the generalizations), and (2) it is not the case that the initial conditions also counterfactually depend on the explanandum (given the generalizations).

Now, are all non-causal explanations asymmetric? I claim that the answer is ‘no’. Consider Euler’s explanation as a case in point. Euler’s theorem supports the *bridges-to-traversability counterfactual*: If all parts of Königsberg were connected to an *even* number of bridges, or if exactly two parts of town were connected to an *odd* number of bridges, then people would not have failed in their attempts to cross all of the bridges exactly once. Thus, Euler’s explanation satisfies the dependency condition of the CTE. However, is Euler’s explanation asymmetric? No, it is not, since Euler’s theorem also supports the *traversability-to-bridges counterfactual*: If people were able to cross all of the bridges exactly once, then all parts of Königsberg would be connected to an even number of bridges, or exactly two parts of town would be connected to an odd number of bridges.

Therefore, Euler’s explanation is symmetric because (1) the traversability counterfactually depends on the number of bridges attached to each part town (given Euler’s theorem), and (2) it is also the case that the number of bridges attached to each part town also counterfactually depends on the traversability (given Euler’s theorem). Hence, according to the CTE there is no explanatory asymmetry in this explanation.

I suspect that not only Euler’s explanation but also other non-causal explanations are not asymmetric, including some, perhaps all, of the examples discussed in this chapter (for a discussion of further examples see Hempel 1965: 352–3; Kistler 2013;

¹⁷ Warning: do not confuse the issue of whether all non-causal explanations are asymmetric with the issue of whether the flagpole-shadow scenario poses a counterexample to the CTE (Reutlinger 2017a: Sect. 5)!

Reutlinger 2017a: section 5). If this is true, then we have an additional reason for classifying those explanations as non-causal. They are non-causal by virtue of not satisfying the Russellian criterion of asymmetry.

6. Conclusion

I have argued for a monist theory of causal and non-causal explanations—the counterfactual theory of explanation. According to the core idea of CTE, causal and non-causal explanations are explanatory by virtue of revealing counterfactual dependencies between the explanandum and the explanans (and by satisfying further conditions). I have argued that the CTE can be successfully applied to five paradigms of non-causal explanations. Using the Russellian strategy, I have justified the claim that these paradigmatic examples are indeed non-causal explanations.

Acknowledgments

I would like to thank Maria Kronfeldner, Marc Lange and Juha Saatsi for charitable and productive feedback.

References

- Achinstein, P. (1983), *The Nature of Explanation* (New York: Oxford University Press).
- Albert, D. (2000), *Time and Chance* (Cambridge, MA: Harvard University Press).
- Andersen, H. (2014), 'A Field Guide to Mechanisms: Part I', *Philosophy Compass* 9: 274–83.
- Andersen, H. (forthcoming), 'Complements, not Competitors: Causal and Mathematical Explanations', *British Journal for the Philosophy of Science*.
- Batterman, R. (2000), 'Multiple Realizability and Universality', *British Journal for the Philosophy of Science* 51: 115–45.
- Batterman, R. (2002), *The Devil in the Details* (New York: Oxford University Press).
- Bennett, J. (2003), *A Philosophical Guide to Conditionals* (Oxford: Oxford University Press).
- Bokulich, A. (2008), 'Can Classical Structures Explain Quantum Phenomena?', *British Journal for the Philosophy of Science* 59: 217–35.
- Cartwright, N. (1989), *Nature's Capacities and Their Measurement* (Oxford: Clarendon Press).
- Elga, A. (2001), 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence', *Philosophy of Science* 68: S313–24.
- Farr, M. and Reutlinger, A. (2013), 'A Relic of a Bygone Age? Causation, Time Symmetry and the Directionality Argument', *Erkenntnis* 78: 215–35.
- Field, H. (2003), 'Causation in a Physical World', in M. Loux and D. Zimmerman (eds.), *The Oxford Handbook of Metaphysics* (Oxford: Oxford University Press), 435–60.
- Fisher, M. (1982), 'Scaling, Universality and Renormalization Group Theory', in F. Hahne (ed.), *Critical Phenomena: Lecture Notes in Physics*, vol. 186 (Berlin: Springer), 1–139.
- Fisher, M. (1998), 'Renormalization Group Theory: Its Basis and Formulation in Statistical Physics', *Reviews of Modern Physics* 70: 653–81.

- Frisch, M. (1998), 'Theories, Models, and Explanation', Dissertation, UC Berkeley.
- Frisch, M. (2014), *Causal Reasoning in Physics* (Cambridge: Cambridge University Press).
- Hempel, C. G. (1965), *Aspects of Scientific Explanation* (New York: Free Press).
- Huneman, P. (2010), 'Topological Explanations and Robustness in Biological Sciences', *Synthese* 177: 213–45.
- Hüttemann, A. (2004), *What's Wrong with Microphysicalism?* (London: Routledge).
- Jansson, L. and Saatsi, J. (forthcoming), 'Explanatory Abstractions', *British Journal for the Philosophy of Science*.
- Kistler, M. (2013), 'The Interventionist Account of Causation and Non-Causal Association Laws', *Erkenntnis* 78: 1–20.
- Ladyman, J. and Ross, D. (2007), *Every Thing Must Go: Metaphysics Naturalized* (Oxford: Oxford University Press).
- Lange, M. (2009), 'Dimensional Explanations', *Noûs* 43: 742–75.
- Lange, M. (2011), 'Conservation Laws in Scientific Explanations: Constraints or Coincidences?', *Philosophy of Science* 78: 333–52.
- Lange, M. (2013a), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Lange, M. (2013b), 'Really Statistical Explanations and Genetic Drift', *Philosophy of Science* 80: 169–88.
- Lange, M. (2016), *Because Without Cause: Non-Causal Explanations in Science and Mathematics* (New York: Oxford University Press).
- Lewis, D. (1973), 'Causation', in D. Lewis, *Philosophical Papers*, vol. II (New York: Oxford University Press, 1986), 159–72.
- Lewis, D. (1979), 'Counterfactual Dependence and Time's Arrow', in D. Lewis, *Philosophical Papers*, vol. II (New York: Oxford University Press, 1986), 32–51.
- Lewis, D. (1986), 'Causal Explanation', in *Philosophical Papers*, vol. II (New York: Oxford University Press), 214–40.
- Lipton, P. (2004), *Inference to the Best Explanation*, 2nd edn. (London: Routledge).
- Loewer, B. (2007), 'Counterfactuals and the Second Law', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (New York: Oxford University Press), 293–326.
- Mach, E. (1872), *Die Geschichte und die Wurzel des Satzes von der Erhaltung der Arbeit* (Prague: Calve).
- Menzies, P. and Price, H. (1993), 'Causation as a Secondary Quality', *British Journal for the Philosophy of Science* 44: 187–203.
- Norton, J. D. (2007), 'Causation as Folk Science', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (New York: Oxford University Press), 11–44.
- Norton, J. D. (2012), 'Approximation and Idealization: Why the Difference Matters', *Philosophy of Science* 79: 207–32.
- Pexton, M. (2014), 'How Dimensional Analysis Can Explain', *Synthese* 191: 2333–51.
- Pincock, C. (2012), *Mathematics and Scientific Representation* (New York: Oxford University Press).
- Pincock, C. (2015), 'Abstract Explanations in Science', *British Journal for Philosophy of Science* 66: 857–82.

- Reutlinger, A. (2013), *A Theory of Causation in the Biological and Social Sciences* (New York: Palgrave Macmillan).
- Reutlinger, A. (2014), 'Why Is There Universal Macro-Behavior? Renormalization Group Explanation as Non-Causal Explanation', *Philosophy of Science* 81: 1157–70.
- Reutlinger, A. (2016), 'Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation', *Philosophy of Science* 83: 733–45.
- Reutlinger, A. (2017a), 'Does the Counterfactual Theory of Explanation Apply to Non-Causal Explanations in Metaphysics?', *European Journal for Philosophy of Science* 7: 239–56.
- Reutlinger, A. (2017b), 'Explanation Beyond Causation? New Directions in the Philosophy of Scientific Explanation', *Philosophy Compass*, Online First, DOI: 10.1111/phc3.12395.
- Reutlinger, A., Hangleiter, D., and Hartmann, S. (2017), 'Understanding (with) Toy Models', *British Journal for the Philosophy of Science*, Online First, <<https://doi.org/10.1093/bjps/axx005>>.
- Rice, C. (2015), 'Moving Beyond Causes: Optimality Models and Scientific Explanation', *Noûs* 49: 589–615.
- Russell, B. (1912/13), 'On the Notion of Cause', *Proceedings of the Aristotelian Society* 13: 1–26.
- Saatsi, J. (2016), 'On Explanations from "Geometry of Motion"', *British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axw007.
- Saatsi, J. and Pexton, M. (2013), 'Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Non-Causal Explanations', *Philosophy of Science* 80: 613–24.
- Saatsi, J. and Reutlinger, A. (forthcoming), 'Taking Reductionism to the Limit: How to Rebut the Anti-Reductionist Argument from Infinite Limits', *Philosophy of Science*.
- Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World* (Princeton: Princeton University Press).
- Salmon, W. (1989), *Four Decades of Scientific Explanation* (Minneapolis: University of Minnesota Press).
- Skow, B. (2014), 'Are There Non-Causal Explanations (of Particular Events)?', *British Journal for the Philosophy of Science* 65: 445–67.
- Strevens, M. (2008), *Depth: An Account of Scientific Explanation* (Cambridge, MA: Harvard University Press).
- van Fraassen, B. (1980), *The Scientific Image* (Oxford: Clarendon Press).
- van Fraassen, B. (1989), *Laws and Symmetries* (Oxford: Oxford University Press).
- von Wright, G. H. (1971), *Explanation and Understanding* (Ithaca: Cornell University Press).
- Weatherall, J. (2011), 'On (Some) Explanations in Physics', *Philosophy of Science* 78: 421–47.
- Wilson, K. (1983), 'The Renormalization Group and Critical Phenomena', *Reviews of Modern Physics* 55: 583–600.
- Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation* (New York: Oxford University Press).
- Woodward, J. (2007), 'Causation with a Human Face', in H. Price and R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited* (New York: Oxford University Press), 66–105.
- Woodward, J. and Hitchcock, C. (2003), 'Explanatory Generalizations, Part I: A Counterfactual Account', *Noûs* 37: 1–24.
- Yourgrau, W. and Mandelstam, S. (1968), *Variational Principles in Dynamics and Quantum Theory* (Philadelphia: Saunders).

5

The Mathematical Route to Causal Understanding

Michael Strevens

1. Introduction

In some scientific explanations, mathematical derivations or proofs appear to be the primary bearers of enlightenment. Is this a case, in science, of “explanation beyond causation”? Might these explanations be causal only in part, or only in an auxiliary way, or not at all? To answer this question, I will examine some well-known examples of explanations that seem to operate largely or wholly through mathematical derivation or proof. I conclude that the mathematical and the causal components of the explanations are complementary rather than rivalrous: the function of the mathematics is to help the explanations’ consumers better grasp relevant aspects of the causal structure that does the explaining, and above all, to better grasp how the structure causally makes a difference to the phenomena to be explained. The explanations are revealed, then, to be causal through and through.

It does not follow that all scientific explanation is causal, but it does follow that one large and interesting collection of scientific explanations that has looked non-causal to many philosophers in fact fits closely with the right kind of causal account of explanation. In that observation lies my contribution to the present volume’s dialectic.

2. Mathematics Gives Us the Gift of Scientific Understanding

Heat a broad, thin layer of oil from below, and in the right circumstances, Rayleigh-Bénard convection begins. At its most picturesque, the convecting fluid breaks up into many hexagonal convection cells, taking on the appearance of a honeycomb. Why that particular shape? An important part of the explanation, it seems, is that the densest possible lattice arrangement of circles in two dimensions is the hexagonal packing: for unrelated reasons, the fluid forms small, circular convection cells; these cells then distribute themselves as densely as possible and fill the interstitial spaces to take on the hexagonal aspect.

The explanation of the honeycomb structure has many parts: the explanation of circular convection cells; the explanation of their tendency to arrange themselves as densely as possible; the explanation of their expanding to fill the interstitial spaces. One essential element among these others is, remarkably, a mathematical theorem, the packing result proved by Lagrange in 1773. To understand the honeycomb structure, then, a grasp of the relevant causal facts is not enough; something mathematical must be apprehended.

* * *

Northern elephant seals have extraordinarily little genetic diversity: for almost every genetic locus that has been examined, there is only one extant allele (that is, only one gene variant that can fit into that genetic “slot”). The reason, as is typical in such cases, is that the seals have recently been forced through a “population bottleneck”. In the late nineteenth century, they were hunted almost to extinction; as the population recovered, it was extremely small for several decades, and in populations of that size, there is a high probability that any perfectly good allele will suffer extinction through simple bad luck—or as evolutionary biologists say, due to random genetic drift.

To explain the genetic homogeneity of contemporary Northern elephant seals, you might in principle construct a real-life seal soap opera, first relating the devastation caused by hunting death after death, and then the rebuilding of the population birth after birth, tracking the fate of individual alleles as the seals clawed their way back to the numbers they enjoy today. But even if such a story should be available—and of course it is not—it would be no more explanatory, and some would say less explanatory, than a suitably rigorous version of the statistical story told above, in which what is cited to explain homogeneity is not births and deaths or even the extinction of individual alleles, but rather the impact of population size on the probability of extinction (and then, not the precise change for any particular allele but just the general trend, with the probability of extinction increasing enormously for sufficiently small populations). The derivation of the fact of this impact takes place entirely within the mathematics of probability theory. Though the explanation also has causal components, it seems to revolve around the mathematical derivation.

* * *

Consider an unusually shaped container—say, a watering can with all openings closed off. Inside the container is a gas, perhaps ordinary air. How does the gas pressure vary throughout the container after the gas is left to “settle down”, that is, after the gas reaches its equilibrium state? The answer is not obvious. Gas pressure is caused by a gas’s molecules pounding on a container’s surfaces. Perhaps the pressure is lower in the neck of the watering can, where there is much less gas to contribute to pressure over the available surface area? Or perhaps it is higher, because at any given moment more of the gas in the can’s neck than in its main body is close to a surface where it can contribute to the pressure?

Assume that at equilibrium, the gas is evenly distributed through the container, so that the density does not vary from place to place, and that the average velocity of gas molecules is the same in each part—a conclusion that it is by no means easy to derive, but the explanation of which I bracket for the sake of this example. Then a short mathematical derivation—essentially, the backbone of the explanation of Boyle’s law—shows that the pressure in the container is the same everywhere. The key to the derivation is that the two factors described above exactly cancel out: there are many more gas molecules in the main section of the watering can, but proportionally more of the molecules in the neck are at any time within striking distance of a surface. The net effect is equal numbers of “strikes” on every part of the can’s—or any container’s—surface. This canceling out is, as in the case of the elephant seals, displayed by way of a mathematical derivation. Mathematics, then, again sits at the center of a scientific explanation.

* * *

An example used to great effect by Pincock (2007) begins with a question about the world of matter and causality: why, setting out on a spring day to traverse the bridges at the center of the city of Königsberg without crossing any bridge twice, would Immanuel Kant fail by sunset to accomplish this task? (The rules governing the attempt to trace what is called an Eulerian path are well known: the path must be continuous and rivers may be crossed only using the bridges in question. You may start and finish anywhere you like, provided that you cross each bridge once and once only.)

The explanation of Kant’s failure is almost purely mathematical: given the configuration of the bridges, it is mathematically impossible to walk an Eulerian path. For any such problem, represent the bridges (or equivalent) as a graph; an Eulerian path exists, Leonhard Euler proved, only if the number of nodes in the graph with an odd number of edges is either two or zero. The graph for the Königsberg problem has four odd-edged nodes.

We could explain Kant’s lack of success by enumerating his travels for the day, showing that no segment of his journey constitutes an Eulerian path. But that explanation seems quite inferior to an explanation that cites Euler’s proof. Perhaps more clearly than in any of the cases described above, this explanation of a material event turns on a mathematical fact, the proof of which is essential to full understanding.

3. The Role of Mathematics in Scientific Explanation

How, then, does mathematics convey understanding of the hexagonal structure of Rayleigh-Bénard convection cells, of the genetic homogeneity of Northern elephant seals, of the uniform pressure of gases at equilibrium regardless of container shape, of the persistent failure of sundry flâneurs’ attempts to traverse, Eulerianly, the bridges of Königsberg? Why, in particular, is it so tempting to say, in each of these cases, that the

phenomenon in question holds *because* of such and such a mathematical fact—that a convection pattern is hexagonal because of Lagrange’s theorem, or that an attempt on the bridges fails because of Euler’s theorem—a locution that seems to place mathematical facts at the heart of certain scientific explanations?

Galileo famously suggested in *The Assayer* that “The Book of Nature is written in mathematical characters”. The Book of Nature is the physical world; this metaphor suggests, then, that mathematics is embedded in nature itself. In that case, perhaps, mathematical properties could explain physical states of affairs by way of mathematical necessitation. I call this notion—it is too nebulous to be called a thesis—the Galilean view of the role of mathematics in explanation.

The Galilean view might be fleshed out in many ways. You might, for example, attribute to abstract mathematical objects—say, the number three—causal powers. Then mathematical necessitation could be understood as a kind of causation, and the examples of mathematical explanation given above as causal. That is not, however, a popular view.

Another possibility runs as follows. Consider a law of nature of the sort usually supposed to describe the effects of causal influence, such as Newton’s second law (never mind that it has been superseded): $F = ma$. The law tells us how an object’s position changes as a consequence of the total impressed force. On the Newtonian worldview, force is doing something in the world: it is making changes in objects’ positions. Both force and position are physical rather than mathematical properties, so there is no mathematical causation here. But could mathematics shape the channels through which causal influence flows? What I have in mind is that there is mathematical structure in the physical world and that causation operates through this structure—in the Newtonian case, for example, perpetrating its effects along lines inscribed in reality by the mathematics of real-numbered second-order differential equations.

If something like this were the case (and of course I have offered only the barest sketch of what that might mean), then it would be no surprise to find mathematics at the core of scientific explanation, dictating the ways in which physical processes may or may not unfold. When we say that a phenomenon obtains *because* of some mathematical fact—say, that no traversal of the bridges can occur *because* of Euler’s theorem—we would mean it literally. It is not that Euler’s theorem is itself a cause (any more than Newton’s second law is “a cause”), but rather that it exhibits a mathematical fact that plays a direct and essential role in the unfolding of the causal processes that constitute attempts at an Eulerian path, a fact that participates in the causal story in a raw and unmediated way, and so whose nature must be grasped by anyone hoping to understand the story.

It is not the strategy of this chapter, however, to defend the causal approach to scientific explanation by upholding a Galilean view. Rather, I will assume a contrary and rather deflationary thesis about the role of mathematics in science—the representational view—and show how, on that view, the examples of mathematically driven scientific explanation cited above ought to be interpreted. I assume the representational

view partly out of an inkling that it may be correct, though I will not argue for such a conclusion here, and partly as a matter of rhetorical strategy, since it allows me to demonstrate that even if, as the representational view implies, there is no prospect whatsoever of mathematical properties playing a role in causation, mathematically driven explanations may nevertheless be understood as wholly causal.

According to the representational view there is either no mathematics in the natural world or mathematics exists in nature in an entirely passive, hence non-explanatory, way. (As an example of the latter possibility, consider the thesis that numbers are sets of sets of physical objects; it follows that they have a physical aspect, but they make up a kind of abstract superstructure that does not participate in the causal and thus the explanatory economy as such.) The role of mathematics in science, and more specifically in explanation, is solely to represent the world's non-mathematical explanatory structure—to represent causes, laws, and the like. A knowledge of mathematics is necessary to understand our human book of science, then, but it is not the content but rather the language that is mathematical. God does not write in mathematical characters—not when she is telling explanatory stories, at least—but we humans, attempting to understand God's ways, represent her great narrative using representational tools that make use of mathematical structures to encode the non-mathematical explanatory facts.

Such a view is suggested by two recent theories of the role of mathematics in science, the mapping account of Pincock (2007) and the inferential account of Bueno and Colyvan (2011). According to both theories, mathematics plays a role in explanation by representing the non-mathematical facts that do the explaining, in particular, facts about causal structure.¹

Can the representational view capture the way in which my example explananda—hexagonal convection cells, elephant seal homozygosity, constant gas pressure—seem to depend on certain mathematical facts? Can they gloss the sense in which the bridges are untraversable *because of* Euler's theorem? The best sense that a representationalist can make of such talk is, I think, that the "because" is figurative: a state of affairs obtains because some non-mathematical fact obtains, and that non-mathematical fact is represented by the mathematical fact, which in a fit of metaphor we proffer as the reason. There is something non-mathematical about the bridges of Königsberg that renders them untraversable; that non-mathematical fact is represented by Euler's theorem and so—eliding, conflating, metonymizing—we say that the failure of any attempt at traversal is "because of" Euler's theorem itself.

If that were all that the representationalist had to say about mathematical explanations in science, this chapter would be short and uneventful. But there is another striking aspect of these explanations besides the "because of", that on the one hand

¹ Other work by Pincock and Colyvan—for example, Pincock (2015)—suggests that these authors may not hold that the mapping and inferential accounts (respectively) exhaust the role of mathematics in science. I take a certain view of the scientific role of mathematics from these authors, then, and to obtain what I call the representational view, I append *And that is all*.

poses a greater challenge to representationalism, and on the other hand to which the representationalist can give a far more interesting reply. This will be my topic for the remainder of the chapter.

The challenge turns from the “because of”, with its apparent implication of an explanatory relation between a mathematical and a physical fact, to the role of mathematical thinking in understanding. When I try to understand what is going on with the bridges or the elephant seals, it seems that thinking mathematically gives me a kind of direct insight into the relevant explanatory structure. Where does that insight come from?

One answer that the representationalist is well positioned to provide is: resemblance. On Pincock’s mapping view, for example, the mathematical structures that feature in explanations are isomorphic to explanatorily relevant structures in the physical world. Grasp the mathematical structure and you grasp the physical structure, at an abstract level at least.

This answer is a good one, but it does not go far enough. Often I gain the majority of my explanatory insight from seeing a mathematical derivation or proof. In many cases, such proofs do relatively little to help me grasp those aspects of mathematical structure that mirror explanatory structure. The isomorphism between the layout of the city of Königsberg and the corresponding graph (Figure 5.1) is obvious. The Euler proof does not make it any clearer—indeed, it simply presupposes it. Much the same can be said for my other paradigm cases.²

The role of proof (or derivation) in these explanations is better described in this way: by following the proof, I see how the mathematical facts necessitate, and so explain, certain physical facts. The Galilean is in a superb position to give this gloss; the representationalist not at all.

To sum up, then, representationalism can go a certain distance in making sense of the role of mathematics in my paradigm explanations. It can to some extent explain away “because” talk, and it can to some extent explain how grasping mathematical structure helps us to grasp explanatory structure. But it does not make very good sense, apparently, of the way in which grasping mathematical proofs helps us to understand physical phenomena. In this respect, the Galilean approach is far superior. That is the challenge to representationalism that I hope to meet in what follows, accounting in representationalist terms for the power of mathematical proof to provide us, in the paradigms above and in other such cases, with causal understanding.

² The treatment in the main text is a little quick, in a way that will become clearer when I present my approach to causal explanation later in this chapter. In the main text, I have taken the aspects of the city layout represented by the Königsberg graph to be the relevant explanatory structure. In fact, the explanatory structure is more abstract than this; it is the fact about the city layout represented by the graph’s having more than two odd-edged nodes. The critique holds, however: whatever the proof does, it goes well beyond helping us to see more clearly that both the city plan and the graph have this property.

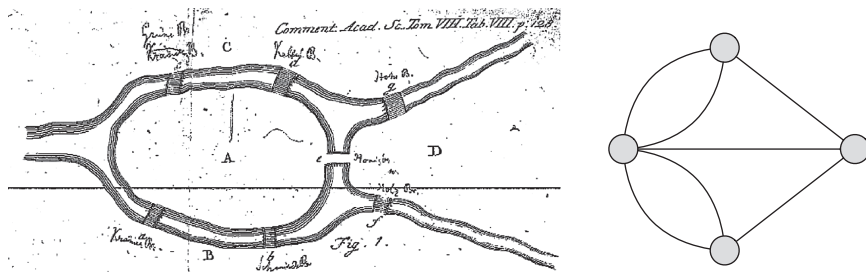


Figure 5.1 Königsberg's bridges.

Left: Euler's sketch of the bridge layout. Right: a graph representing the bridges.

(Source: MAA Euler Archive, <<http://eulerarchive.maa.org>>)

4. Explanatory Relevance as Causal Difference-Making

Scientific explanation, according to the approach I will adopt in this chapter without argument, is a matter of finding causal difference-makers. Much of what I want to say could be framed in terms of any sophisticated difference-making account, but I will—naturally—rely here on my own “kairetic account” (Strevens 2004, 2008).

The raw material of explanation, according to the kairetic account, is a fundamental-level structure of causal influence revealed by physics. For simplicity's sake, assume that we live in a classical world constituted entirely of fundamental-level particles that interact by way of Newtonian forces. Then the fundamental-level causal structure is the network of force, that is, the totality of forces exerted, at each instant of time, by particles on one another, whether gravitational, electromagnetic, or something else. This web of force, together with other relevant facts about the particles—their positions, their velocities, their inertial masses, and so on—determines each particle's movements, and so determines everything that happens in the material world. That, at least, is the Newtonian picture, which I assume here; modern physics of course requires some revisions.³

The web of force is vast and dense, titanic and tangled; it is beyond the power of human science to represent any significant part of it explicitly and exactly. Were scientific explanation to require us to provide an exhaustive inventory of the forces acting on a particle at any given time—an inventory that would include, for a particle of non-zero mass, the gravitational influence of every other massive particle in the universe—we would have no prospect whatsoever of constructing complete explanations.

We aspiring explainers avoid such impossible demands, because we are interested in explaining mainly high-level events and states of affairs, and explanation requires that

³ I will not countenance the possibility that modern physics will show the world to be devoid of causality, or the milder but still alarming possibility that causality might only “emerge” at levels higher than that of fundamental physics.

we identify only the aspects of the causal web that make a difference to whether or not those events occurred or those states of affairs obtained—which difference-makers are far more sparse than causal influences.

To fill out this picture, consider event explanation in particular. With a rebel yell, Sylvie hurls a cannonball at the legislature's prize stained-glass window; it shatters. What explains the shattering? In asking this question, I am interested in why the window shattered rather than not shattering. The explainers I have in mind are the ball's hitting the window, Sylvie's throwing the ball, the window's composition—and not much more. I could have asked a different question: why did the window shatter in exactly the way that it did, with this shard traveling in this direction at this velocity and so on? To answer such a question I would have to take into account many more causal influences—many more Newtonian forces—that acted on the shattering. Sylvie's yell, for example, caused the window to vibrate a little, which accounts in part for the exact trajectories of the myriad shards.

The contrast between these two questions—the question of why the window broke, and the question of why the window broke in precisely the way that it did—illustrates the difference between a high-level event such as the breaking and the low-level or “concrete” event that realizes the breaking, that is, the window's breaking in precisely such and such a manner, specified down to the most minute details of each molecule's trajectory. Because explanation is about finding difference-makers, an answer to the latter question must cite pretty much every causal influence on the window, while an answer to the former question ignores elements of the causal story whose only impact is on *how* the window broke, and focuses instead on those elements that made a difference to *whether* the window broke. Sylvie's insurrectionary cry made a difference to the precise realization of the window's shattering, and helps to explain that concrete event, but it made no difference to whether not the window shattered; it thus plays no part in explaining the high-level event of the shattering.

Science's explanatory agenda is focused almost exclusively on high-level events as opposed to their concrete realizers. Biologists want to explain why humans evolved large brains, but they are not (on the whole) interested in accounting for the appearance of every last milligram of brain tissue, except insofar as it casts light on the bigger question. Planetary scientists would like to explain the formation of the solar system, but they certainly have no interest in explaining the ultimate resting place of individual pebbles. Economists are interested in explaining why the recent financial crisis occurred, but they are not (on the whole) interested in explaining the exact dollar amount of Lehman Brothers' liabilities. In each case, then, the would-be explainers must decide which elements of the causal web, the densely reticulated network of influence responsible for all physical change, were significant enough to make a difference to whether or not the phenomena of interest occurred—to the fact that human brains grew, that the solar system took on its characteristic configuration, that between 2007 and 2008 the global financial system warped and fractured.

The role of a theory of explanation is to provide a criterion for difference-making that captures this practice—that classifies as difference-makers just those aspects of the causal web that are counted as such by scientific explainers. An obvious choice is a simple counterfactual criterion: a causal influence on an event is an explanatory difference-maker for the event just in case, had it not been present, the event would not have occurred. Had Sylvie thrown the cannonball without a word, the window would still have broken, so her vocal accompaniment is not a difference-maker for the breaking. But had she not thrown the cannonball at all, the window would have remained intact; thus, her throwing is a difference-maker. As is well known from the literature on singular causation, however, the counterfactual criterion fails to capture our judgments of difference-making, both everyday and scientific: Bruno might have been standing by to break the window in case Sylvie failed; in these circumstances, it is no longer true that had Sylvie refrained from throwing, the window would not have broken (Lewis 1973). The counterfactual criterion counts her throw as a non-difference-maker in Bruno's presence, but we want to say that, since Sylvie did in fact throw the ball and her ball broke the window, her throw was a decisive difference-maker for the breaking.

In the light of these and other problems for the counterfactual approach (Strevens 2008: chapter 2), I have proposed an alternative criterion for difference-making. The “kairetic criterion” begins by supposing the existence of a complete representation of the relevant parts of the causal web, that is, a complete representation of the causal influences on the event to be explained. This representation takes the form of a deductive argument in which effects are deduced from their causes along with causal laws. In the case of the window, for example, the trajectory of each shard of glass will be deduced from the relevant physical laws and initial conditions—the trajectory and makeup of the incoming cannonball, the molecular constitution and structure of the window and its connection to its frame, and all other relevant environmental circumstances, including in principle the gravitational influence of the distant stars. Such a deduction shows how the breaking in all of its particularity came about as the aggregate result of innumerable causal influences; it is a representation of the complete causal history of the breaking.

A few comments on this canonical representation of the causal process leading to the breaking. First, it is of course quite beyond the powers of real scientists, even very well-funded and determined real scientists, to construct such a representation. The canonical representation's role is to help to lay down a definitive criterion for causal difference-making; in practice scientists will decide what is likely to satisfy the criterion using a range of far more tractable heuristics. To give a simple example, the gravitational influence of other stars almost never makes a difference to medium-sized terrestrial events such as window breakings; the stars can therefore from the outset be ignored.

Second, there is much more to say about the structure in virtue of which the canonical representation represents a causal process. I have said some of it in Strevens (2008), chapter 3; in the current chapter, however, the details are of little importance. I will simply assume that there is some set of conditions in virtue of which a sound deductive

argument represents a causal process or, as I will say, in virtue of which it qualifies as a veridical causal model.

Third, in assuming that the explanandum can be deduced from its causal antecedents, I am supposing that the process in question is deterministic. In the stochastic case, what is wanted is rather the deduction of the event's probability, as suggested by Railton (1978). Again, I put aside the details; for expository purposes, then, assume determinism.

On with the determination of difference-makers. The idea behind the kairetic account is simple: remove as much detail as you can from the canonical representation without breaking it, that is, without doing something that makes it no longer a veridical causal model for the event to be explained. The "removal" consists in replacing descriptions of pieces of the causal web with other descriptions that are strictly more abstract, in the sense that they are entailed by (without entailing) the descriptions they replace and that they describe the same subject matter or a subset of that subject matter.⁴

In the case of the broken window, for example, much of the structure of the cannonball can be summarized without undermining the veridicality or the causality of the canonical model. What matters for the deduction is that the ball has a certain approximate mass, size, speed, and hardness. The molecule-by-molecule specification of the ball's makeup that appears in the canonical representation can be replaced, then, by something that takes up only a few sentences. Likewise, the fact of Sylvie's war cry can be removed altogether, by replacing the exact specification of her vocalization with a blanket statement that all ambient sound was within a certain broad range (a range that includes almost any ordinary noises but excludes potential window-breakers such as sonic booms).

When this process of abstraction has proceeded as far as possible, what is left is a description of the causal process leading to the explanandum that says as little about the process as possible, while still comprising a veridical causal model for the event's production. The properties of the process spelled out by such a description are difference-making properties—they are difference-makers for the event. The approximate mass, size, speed, and hardness of the cannonball make a difference to the window's breaking, then, but further details about the ball do not. Nothing about Sylvie's yell makes a difference except its not exceeding a certain threshold. These difference-makers are what explain the window's breaking; aspects of the causal web that do not make a difference in this sense, though they may have affected the event to be explained—determining that this shard went here, that one there—are explanatorily irrelevant.

Observe that the kairetic account envisages two kinds of causal relation. The first kind is causal influence, which is revealed by the correct fundamental-level theory

⁴ The removal operation is constrained additionally by a requirement that the representation should remain "cohesive", which ensures that abstraction does not proceed by adding arbitrary disjuncts. Cohesion is relevant to some aspects of the following discussion, but for reasons of length I will put it aside.

of the world and serves as the raw material of causal explanation. The second is causal difference-making, an explanatory relation that links various properties of the web of influence to high-level events and other explananda. Difference-making relations are built from causal influence according to a specification that varies with the phenomenon to be explained.

The cases of mathematically driven understanding presented above, you will note, involve high-level difference-making, in which many *prima facie* causally significant features of the setup turn out not to be difference-makers: the development of particular convection cells, the shape of particular containers, the twists and turns taken in an attempt to travel an Eulerian path around Königsberg. That is an important clue to what mathematics is doing for us, as you will shortly come to see.

My goal is to show that mathematically driven explanations in science are causal, in the manner prescribed by the kairetic or some other difference-making account. My working assumption is that the role of mathematics in science, including explanation, is purely representational, standing in for inherently non-mathematical features of nature. If mathematics is an aid to scientific explanation, then, its assistance had better be indirect, arriving in virtue of something that it does as a representer of causal structure (though not necessarily representation *simpliciter*). To see what that something might be, I turn to the topic of understanding.

5. Mathematics and Causal Understanding

According to what I have elsewhere dubbed the “simple view” of understanding, to understand a phenomenon is to grasp a correct explanation for that phenomenon (Strevens 2013). Combining the simple view with the kairetic account of scientific explanation yields the following thesis: to understand a material event is to grasp the difference-making structure in which that event is embedded and in virtue of which it occurred. The explanation, on this approach, is “out there”: it is a collection of causal facts—causal difference-making relations, to be precise—waiting to be discovered by science. Understanding is the cognitive achievement realized by epistemically connecting to the explanation, to these facts, in the right way.

The philosophy of understanding is much concerned with what counts as the “right way”. Is it a matter of having deep knowledge of the relevant domain or is it rather a matter of possessing some ability that goes beyond mere knowledge? In this chapter I will keep my distance from these debates, assuming that at a minimum, in order to grasp a causal difference-making structure a seeker of understanding must grasp both the nature of the difference-makers for the explanandum and the way in which they make the difference that they do.

What role can mathematics play in all of this? Difference-making structures are not inherently mathematical—that, at least, is my representationalist working assumption—but mathematics might nevertheless help us to get a grip on such a structure, attaining

the kind of epistemic connection to the difference-makers and their difference-making that constitutes understanding. Here's how.

A part of grasping an explanation is to apprehend clearly the topology of the relevant relations of causal difference-making. Mathematics is often central to this task, transparently and concisely representing the structure to be grasped, whether by way of a directed graph (the formal equivalent of boxes and arrows), a set of differential equations, a stochastic dynamical equation, or in some other way.

In performing this function, mathematics does just what the representationalist claims: it provides compact, precise, in many cases tailor-made symbolic systems to represent relationships in the world. As I argued in section 2, however, this representational role, vital though it may be, does not cast much light on the importance of mathematical derivation or proof in causal understanding. A system of definitions seems to be sufficient to undergird a system of representation; theorems derived from those definitions add no representational power and in many cases make the representation no more effective than it was before. The graph representing the bridges of Königsberg presents the essential structure of the problem just as plainly and perspicuously to someone who does not know of Euler's proof as to someone who does. Yet understanding the proof seems absolutely central to understanding why Kant failed to complete an Euler walk around the bridges on some fine day in May.

To appreciate the function of proof, we need to turn to another facet of causal understanding. Knowing the difference-makers is not enough, I submit, to grasp an explanation; you must understand *why* they are difference-makers, or in other words, you must grasp the reasons for their difference-making status.

Go back to the cannonball through the window, to begin with a very simple case. What does it take to understand why the window shattered? You must, at the very least, grasp the fact that the ball's hitting the window caused it to shatter—that the striking was a causal difference-maker for the shattering. But there is more to explanation and understanding than this. It is also important to see in virtue of what aspects of the situation the ball caused the shattering, that is, to see how and why it was a difference-maker. In part this is a matter of grasping (at the appropriate level of abstraction) the structure of the underlying causal process: the transfer of momentum to parts of the window; the stress thereby placed on the bonds holding the window together; the catastrophic failure of the bonds due to their inelasticity.

Equally, it is a matter of seeing that these elements of the causal web were sufficient in themselves to bring about the breaking, that they and nothing else (aside from their own causes, such as Sylvie's throwing) were the difference-makers for the breaking. This insight comes most directly and also most deeply through an application of the kairetic criterion, that is, through seeing that it is possible to abstract away from all other properties of the web while still deriving the fact of the shattering. And mathematical proof is the royal road to this goal: a proof, once fully understood, shows us with unrivaled immediacy what is and is not required for a derivation. In so doing, it shows

us why difference-makers satisfy the criterion for difference-making—why they are difference-makers.

The proof, in short, because it is not part of the difference-making structure, is not a part of the explanation. Its role is not to explain but to help us to grasp what explains—to see the difference-makers for what they are—and so to help us to understand.

* * *

Let me now return to the examples of mathematically driven understanding that I presented above: hexagonal Rayleigh-Bénard convection cells, genetic uniformity in elephant seals, the irrelevance of container shape to gas pressure, and the bridges of Königsberg. In each of these cases, I suggest, the value of mathematical proof lies in its helping us to grasp which aspects of the great causal web are difference-makers for the relevant explanandum and why—and complementarily, helping us to grasp which aspects of the web are not difference-makers and why. What makes these particular examples especially striking, and the underlying mathematical proofs especially valuable, is that there are many important-looking parts of the causal story that turn out, perhaps contrary to initial expectations, to be non-difference-makers. The mathematics shows us why, in spite of their substantial causal footprint, they make no difference in the end to the phenomenon to be understood.

Consider the elephant seals. Large numbers of seal alleles went extinct in a short time, but the extinction had nothing to do with the intrinsic nature or developmental role of those alleles. They simply suffered from bad luck—and given the small size of the seal population in the early twentieth century, it was almost inevitable that bad luck would strike again and again, eviscerating the gene pool even if the species as a whole endured. The mathematics reveals, then, that the extinction of so many alleles was due to a haphazard mix of causal processes—mostly to do with mating and sex (though also including death by accident and disease)—whose usual aleatory effect on the makeup of the gene pool was powerfully amplified by the small size of the population, wiping out almost all the elephant seals' genetic diversity.

The mass extinction of seal alleles has a causal explanation, then—a highly selective description of the operation of the relevant part of the causal web, that is, the ecology of the Northern elephant seal over several decades. To see that this is the correct explanation, however—to see that in spite of its high level of abstraction, its omission of so much that seems important, it contains all the explanatorily relevant factors, all the difference-makers—mathematical thinking is invaluable. It is the mathematics that enables you to see both how cited factors such as mating choice and sex, not normally regarded as indiscriminate extinguishers of biological diversity, erased so many alleles, and why as a consequence many uncited factors better known for their selective power, above all the various genes' phenotypic consequences, were not difference-makers at all.

Or consider gas pressure. The essence of the explanation for a gas's uniform pressure on all surfaces of its container is causal; it embraces both the causal process by which

the gas spreads itself evenly throughout the container, creating a uniform density, and the process by which a gas in a state of uniform density creates the same pressure on all surfaces. As in the elephant seal case, however, the explanation has very little to say about these causal processes. It barely mentions the physics of molecular collision at all, and the container walls themselves figure in the story only in the most abstract way. The walls' shape, in particular—the geometry of the container as a whole—is conspicuous only by its omission from the explanation. Mathematics helps us to grasp this explanation by showing us why the details of collision and container shape make no difference—in effect, by showing us that uniform pressure can be derived from a description of a few abstract properties of the gas however the details are filled out.

* * *

Now let me tackle the tantalizing Königsberg case. Here, it is tempting to say, mathematics takes over from causal explanation altogether, yielding a bona fide example of the explanation of a physical fact—Kant's failure to complete an Euler walk around the Königsberg bridges on May Day, 1781—that lies entirely beyond causation. I assimilate it, nevertheless, to the other examples in this chapter. The explanation of Kant's failure takes the form of a highly abstract description of the relevant piece of the causal web—that is, of his day's wanderings—that extracts just the difference-making features of the web. The role of the mathematics is not strictly speaking explanatory at all; rather, it helps us to understand why a certain ultra-abstract description of Kant's movements that day constitutes a correct explanation, that is, a description which includes all the difference-makers and therefore omits only those properties of the web that made no difference to the event to be explained.

To see this, start with a different bridge-traversal task: say, the task of visiting each of the four Königsberg land masses (two islands and the two banks of the river) exactly once—or in more abstract terms, the task of visiting each node in the corresponding graph exactly once, which in graph theory is called a Hamiltonian walk. Such a journey is possible in the Königsberg setup, but it is also possible to go wrong, choosing to traverse a bridge that takes you back to a landmass you have already visited before the walk is complete. Suppose that Kant attempts a Hamiltonian walk. He chooses a good starting point (in this case, all starting points are equally good); he travels to another node (so far, so good); but then he makes a bad decision and travels back to his starting point without visiting the other two nodes in the graph. Why did his attempt fail? He made a wrong turn. A brief explanation would simply lay out the facts that make the turn a bad one and then note that he made it nevertheless. The same is true for the case where he fails because he chooses a bad starting point, say the middle node in the graph shown in Figure 5.2.

A great deal is left out of these explanations. They omit everything about Königsberg except the barest facts as to the layout of its bridges and everything about Kant's means of locomotion that is not relevant to his conforming to the rules for making a graph-theoretic walk. Also omitted, most importantly, is any specification of Kant's travels



Figure 5.2 A Hamiltonian walk.

To complete a Hamiltonian walk of this graph, begin at one end or the other but not in the middle.

after the point at which he makes a bad decision (either choosing a wrong turn or a wrong starting point). If a fatal error has already been committed, these facts make no difference to his failing to complete a Hamiltonian walk, because they can be deleted from the causal story without undermining its entailment of failure.

In the case of a bad choice of starting point, then, there is no description at all of the movement from land mass to land mass (that is, from node to node); the explanation is over almost as soon as it begins, with the description of the problem, the initial bad choice of starting point, and a certain fact about the bridges: from that starting point, no Hamiltonian path can be traced. Yet, I claim, like any causal difference-making explanation, this one is a description of the relevant causal process in its entirety. It does not describe everything about that process—it leaves out the non-difference-making properties—but what it describes is present in the explanation only because it is a feature of the causal process.

Indeed, in its omission of any aspect of Kant's route after the initial choice of starting point, the explanation is not so different from, say, the explanation of genetic homogeneity in elephant seals. There, too, there is no attempt to trace a particular causal trajectory. What matters instead is a rather abstract feature of the process, that it contains many events that act like random samplers of genes, and that the intensity of the sampling is such as to very likely exclude, over a certain length of time, almost every allele from the gene pool. Likewise, what matters about Kant's walk is that it is a journey carried out under a certain set of constraints (formally equivalent to a walk around a graph), that it began from a certain point, and that under these constraints, no journey beginning from that point can complete a Hamiltonian walk. The actual route taken is not a difference-maker.

From there, it is one short step to the explanation of Kant's inability to complete an Euler walk: here all possible starting points are "bad", so the identity of Kant's actual starting point is also not a difference-maker. What is left in the explanation is only generic information: the structure of the bridges and land masses and the aspects of Kant's journeying that make it formally equivalent to a walk around a graph. It is a description of a causal process—a description adequate to entail that the causal process ended the way it did, in Euler-walk failure—yet it has nothing to say about the specifics of the process, because none of those specifics is a causal difference-maker. Euler's theorem helps you to understand why.⁵

The case is very similar to another well-known example in the philosophy of explanation first brought into the conversation by Sober (1983) and then discussed

⁵ Note that the most general version of the theorem is needed to determine correctly all the difference-makers. Consider a weaker version (of no mathematical interest) that applies only to systems with an odd number of bridges. Armed only with such a theorem, you would be unable to grasp the non-difference-making status of the fact that the number of bridges is odd.

extensively by (among others) Strevens (2008), namely, the explanation why a ball released on the inside lip of an ordinary hemispherical salad bowl will end up, not too long later, sitting motionless at the bottom of the bowl. The explanation identifies certain important features of the relevant causal web, that is, of the causal process by which the ball finds its way to the bowl's bottom: the downwardly directed gravitational field, the convex shape of the bowl, the features in virtue of which the ball loses energy as it rolls around. But it has nothing to say about the ball's actual route to the bottom—nothing about the starting point (that is, the point on the rim of the bowl where the ball was released), nothing specific about the manner of release, and nothing about the path traced in the course of the ball's coming to rest at the foreordained point.

The only philosophically important difference between the ball/bowl explanation and the bridges explanation is that mathematics plays a far more important role in helping us to grasp why the specified properties of the bridges setup are difference-makers and the omitted properties are not. In the case of the bowl, simple physical intuition makes manifest the irrelevance of the release point and subsequent route; in the case of the bridges, we need Euler's proof to see why Kant's choice of route makes no difference to the end result.

To sum up: ordinary causal explanations such as the cannonball and the window, equilibrium explanations such as the ball in the bowl, statistical explanations such as elephant seal homozygosity and uniform gaseous pressure, and what some have taken to be purely mathematical explanations such as the famous Königsberg bridges case, are all descriptions of the causal processes leading to their respective explananda, couched at a level of description where only difference-makers appear in the explanatory story. Sometimes the difference-makers entail that the system takes a particular causal trajectory, but often not—often the trajectory is specified only at a very qualitative or diffuse level, and sometimes not at all.

Mathematics has more than one role to play in the practice of explaining, but its truly marvelous uses tend to involve the application of theorems to demonstrate the explanatory power—the difference-making power—of certain abstract properties of the causal web, and even more so the lack of difference-making power of other salient properties of the web. Deployed in this way, the mathematics is not a part of the difference-making structure itself; nor does it represent that structure. Rather, it illuminates the fact that it is this structure rather than some other that makes the difference; it allows us to grasp the reasons for difference-making and non-difference-making, so bringing us epistemically closer to the explanatory facts—and thus making a contribution, if not to explanatory structure itself, then to our grasp of that structure and so to our understanding of the phenomenon to be explained.

6. Explanation Beyond Causation?

What lessons can be drawn about causal explanation? Does the spectacular use of mathematics in cases such as the elephant seals or the Königsberg bridges show that scientific explanation goes beyond causation? Even if everything I have said so far

is correct, it might be maintained that the Königsberg explanation, though it has causal content, is too abstract to constitute a causal explanation. Let me consider, and repudiate, some arguments to that effect.

I begin with a recapitulation. My view that the Königsberg explanation is a causal explanation is not based on the weak and inconclusive observation that the explanatory model has some causal content. It is based on the observation that the model's sole purpose is to pick out the properties of the web of causal influence that, by acting causally, made a difference to whether or not the explanandum occurred. The model is, in other words, exclusively concerned with detailing all relevant facets of the causation of the phenomenon to be explained. It aims to do that and nothing else. If that's not a causal explanation, what is?

Objection number one: a genuine causal explanation not only lays out the causal difference-makers but also tracks the underlying causal process, whether it is a stroll around Königsberg or the trajectory taken by a ball on its way to the bottom of a salad bowl. Classify all scientific explanations, then, into two discrete categories, tracking and non-tracking. The tracking explanations not only cite causal structure but also show how this structure guides an object or a system along a particular path that constitutes or results in the occurrence of the explanandum. The non-tracking explanations may cite causal structure, but they get to their explanatory endpoints not along specific paths but by other means, such as a demonstration that the endpoint is inevitable whatever path is taken. The non-tracking explanations are (according to the objection) non-causal.⁶

Such an explanatory dichotomy is, I think, indefensible. There is an enormous range of causal explanations saying more and less in various ways about the underlying causal web. The dimensions of abstraction are many, and explanations pack the space, forming a continuum of abstraction running from blow-by-blow causal tales that run their course like toppling dominoes to magical equilibrium explanations that pull the explanandum out of the causal hat in a single, utterly non-narrative, barely temporal move—and with, perhaps, a mathematical flourish. Sometimes an explanation begins narratively, like the explanation of Kant's failure to trace a Hamiltonian path that begins with his bad decision as to a starting point, only to end quite non-narratively, with a proof that from that point on, failure was inevitable. Or it might be the other way around (if, say, the choice of starting point doesn't matter but later decisions do). Further, there are many degrees of abstraction on the way from simple narrative to magic hat. The elephant seal explanation tells a causal story of relentless extinction by random sampling, but the extinctions are characterized only at the most typological level. The gas pressure explanation is quite viscerally causal on the one hand—molecules colliding with one another and pounding on the walls of their container—yet on the other hand extraordinarily abstract, compressing heptillions of physical parameters,

⁶ To make such a case for the non-causality of equilibrium explanations was Sober's aim in introducing the "ball in the bowl".

the positions and velocities of each of those molecules, into a few statistical aggregates. And these are only a handful of the possible routes to abstraction, each one tailor-made for a particular explanandum.

Consequently, I see no prospect whatsoever for a clear dividing line between causal tracking explanations and non-causal non-tracking explanations. The gulf between a conventional causal narrative and the Königsberg explanation is vast. But it ought not to be characterized as one of causal versus non-causal character, in part because that is to suppose a dichotomy where there is a continuum of abstraction and in part because everywhere along the continuum the aim of explanation is the same: to find whatever properties of the causal web made a difference to the explanandum.

Objection number two draws the line between causal and non-causal descriptions of the web of influence in a different place, with fewer explanations on the non-causal side. The Königsberg explanation (observes the objector) is special even among very high-level, very abstract causal explanations: it deals in mathematical impossibility rather than physical or nomological impossibility. Does that difference in the guiding modality not constitute a discontinuity?

To put it another way, failure to complete an Euler walk of the Königsberg bridges is inevitable not only in universes that share our world's laws of nature. If our physics were Newtonian, Kant could not complete the walk. Even if it were Aristotelian, he could not complete the walk. Were Kant descended from lizards rather than apes, he could not complete the walk; likewise if he were a silicon-based rather than a carbon-based life form. The implementation of his psychology is equally beside the point: whether plotting his turns with neural matter, with digital processing, or using the immaterial thought stuff posited by dualist philosophers, he would be unable to pull off an Euler walk, for the very same reason in each case.

The explanation of Kant's failure, then, has enormous scope: it applies to many possible worlds other than our own provided that a few simple posits hold—namely, that the network of bridges has a certain structure and that the Kantian counterpart's movements are constrained so as to conform to the rules defining a graph-theoretic walk (movement is always from one node to another neighboring node along an arc).

Does that make the Königsberg explanation *sui generis*? It does not. Any explanatory model that abstracts to some degree from the fundamental physical laws accounts for its explanandum not only in the actual world but also in worlds whose laws differ from the actual laws solely with respect to features from which the model abstracts away. Since almost all explanatory models are abstract not only in what they say about particulars but also in what they say about the laws in virtue of which the particulars are causally connected, almost all explanatory models have a modal extent that reaches beyond the nomologically possible. The more they abstract, the wider the reach.

The Newtonian model for the cannonball's breaking the window, for example, abstracts from the exact value of the gravitational constant, implying a shattering for any value in the vicinity of the actual value—any value not so high that the cannonball thuds to the ground before it gets to the window or so low that it overshoots the window.

The model thus applies to a range of broadly Newtonian theories of physics, differing in the value they assign to the constant.

More interestingly, I suggest that the simple kinetic theory of gases gives valid explanations in both classical and quantum worlds, and that the elephant seal explanation is valid for a great variety of possible biologies that depart considerably from the way things work here on Earth, in both cases because the explanatory models assume rather little about the physical underpinnings of the processes they describe. The great modal reach of the Königsberg model is, then, far from unusual. It is an exceptional case because it calls for so high a level of explanatory abstraction, but its specialness is a matter of degree rather than of kind.

My response to both the second and the first objections, then, is to argue for a continuum (practically speaking, at least) of explanatory models in every relevant dimension, and to reject any attempt to draw a meaningful line across this continuum as invidious. Marc Lange (2013) has recently suggested a variant on the second objection that attempts to find a non-arbitrary line founded in gradations of nomic necessity.

The explanandum in question is that a double pendulum has at least four equilibrium configurations. Lange offers an explanation in the framework of Newtonian physics that he takes to be non-causal. The explanation depends on the fact that all force laws must conform to Newton's second law ($F = ma$) but on no further facts about the laws in virtue of which the pendulum experiences forces. Writing that "although these individual force laws are matters of natural necessity, Newton's second law is more necessary even than they", Lange suggests drawing the line between causal and non-causal explanations at the point that separates the force laws' physical necessity on the one hand, and the second law's higher grade of nomological necessity on the other. An explanation that depends only on this higher grade (or a grade higher still) is, he holds, non-causal. (Lange calls such explanations "distinctively mathematical", but that strikes me as a misnomer: $F = ma$ is no more mathematical than $F = GMm/r^2$; the higher necessity of $F = ma$ is nomological rather than mathematical necessity.)

Lange's view hinges on the proposition that there is something special about the line between the individual force laws and the second law. But what? It is not simply that the second law is more necessary: as I have shown above, in the space of valid scientific explanations, there is a continuum of modal strength running all the way from very particular contingent facts, to very particular facts about the actual laws of nature, to rather more abstract facts about the actual laws, and so on up to very abstract properties such as those that underwrite the kinetic theory in both classical and quantum worlds.

Why, then, is this the particular line in modal space at which the causality "goes away"? Lange tells us, writing of the double pendulum explanation (2013: 19):

This is a non-causal explanation because it does not work by describing some aspect of the world's network of causal relations.... Newton's second law describes merely the framework within which any force must act; it does not describe (even abstractly) the particular forces acting on a given situation.

This, I think, is false. Newton's second law does describe, very abstractly, a property of the particular forces (and force laws): it says that they conform to Newton's second law. That is a fact about them. More generally, that a causal law operates (of necessity or otherwise) within a particular framework is a fact about that law. Thus it is a fact about the world's network of causal relations.

Two further remarks about Lange's view. First, it is inspired by a metaphysics in which there are laws at different modal strata: say, force laws at the bottom stratum and then constraints on force laws, such as Newton's second law, at a higher stratum. The laws at each stratum impose non-causal constraints on the stratum below, while the laws at the bottom stratum are causal laws that determine the course of events in the natural world. Lange would say that the higher-level laws are not acting causally; I say that their action on the bottom-level laws is not causal, but their action on events most certainly—albeit indirectly—is.

Second, Lange treats the Königsberg bridges in a similar way to the double pendulum case (if only in passing). In the Königsberg case, however, the higher and therefore putatively non-causal grade of necessity is not a kind of nomological necessity; it is mathematical necessity. This picture is, I think, incompatible with representationalism, on which mathematics has no power to constrain what laws there can be. (The representationalist holds that our representations of the laws must conform to mathematical principles because the principles are built into our system of representation, not because they are built into the world.) I have assumed rather than argued for representationalism, so this cannot be regarded as a refutation of Lange's treatment of the bridges, but it does put his strategy outside the scope of this chapter.

* * *

Is all scientific explanation causal? I have not argued for such a sweeping conclusion; what I have done is to remove an obstacle to maintaining such a view, and to argue more generally against any attempt to draw a line distinguishing "non-causal" from causal descriptions of the causal web.

Let me conclude by noting that there is an entirely different way that non-causal explanation might find its way into science: some scientific explanations might be constructed from non-causal raw material, say, from a kind of non-directional nomological dependence rather than causal influence. Such explanations would describe difference-making aspects of the web of acausal nomological dependence; they would be non-causal from the bottom up. But whether there are any such things is a topic for another time.

References

- Bueno, O. and Colyvan, M. (2011), 'An Inferential Conception of the Application of Mathematics', *Noûs* 45: 345–74.
- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.

- Lewis, D. (1973), 'Causation', *Journal of Philosophy* 70: 556–67.
- Pincock, C. (2007), 'A Role for Mathematics in the Physical Sciences', *Noûs* 41: 253–75.
- Pincock, C. (2015), 'Abstract Explanations in Science', *British Journal for the Philosophy of Science* 66: 857–82.
- Railton, P. (1978), 'A Deductive-Nomological Model of Probabilistic Explanation', *Philosophy of Science* 45: 206–26.
- Sober, E. (1983), 'Equilibrium Explanation', *Philosophical Studies* 43: 201–10.
- Strevens, M. (2004), 'The Causal and Unification Approaches to Explanation Unified—Causally', *Noûs* 38: 154–76.
- Strevens, M. (2008), *Depth: An Account of Scientific Explanation* (Cambridge, MA: Harvard University Press).
- Strevens, M. (2013), 'No Understanding without Explanation', *Studies in History and Philosophy of Science* 44: 510–15.

6

Some Varieties of Non-Causal Explanation

James Woodward

1. Introduction

The topic of non-causal explanation is very much in vogue in contemporary philosophy of science, as evidenced both by this volume and by many other recent books and papers. Here I explore some possible forms of non-causal scientific explanation. The strategy I follow is to begin with the interventionist account of causal explanation I have defended elsewhere (Woodward 2003) and then consider various ways in which the requirements in that account might be changed or loosened to cover various putative non-causal explanations. I proceed in this way for a variety of reasons. First, causal explanations are generally regarded as at least one paradigm of successful explanation, even if there is disagreement about how such explanations work and what sorts of features mark them off as causal. A general account of explanation that entailed that causal claims were never explanatory or that cast no light on why such claims are explanatory is, in my opinion, a non-starter. Moreover, although it is possible in principle that causal and non-causal explanations have no interesting features in common, the contrary assumption seems a more natural starting point and this also suggests beginning with causal explanations. Second, if one is going to talk about “non-causal” explanation, one needs a clear and well-motivated notion of causal explanation to contrast it with. Third, we have a fairly good grasp, in many respects of the notion of causation, and how this connects to other concepts and principles that figure in science. These include connections to probability, as expressed in, e.g., the principle of the common cause and the Causal Markov condition and, relatedly, connections between causal independence and factorizability conditions, as described in Woodward (2016b). Also of central importance is the connection between causal claims and actual or hypothetical manipulations or interventions, as described in Woodward (2003). Within physics, notions of causal propagation and process, where applicable, are connected to (and expressed in terms of) other physical claims of various sorts—no signaling results in quantum field theory, prohibitions on space-like causal connections, and so on. To a

considerable extent, we lack corresponding connections and constraints in connection with non-causal forms of explanation. This is not a good reason for neglecting the latter, but again suggests a strategy of using what we understand best as a point of departure.

Finally, another important point about the contrast between causal and non-causal explanations: It is tempting to suppose not just that these are different (which is a pre-supposition of any discussion of this topic) but that there are scientific theories that exclusively provide one rather than the other; in other words, that there are non-causal explanations that proceed independently of any variety of causal explanation (or independently of any sort of causal information) and perhaps conversely. It seems to me that the truth is often more complicated and nuanced; often plausible candidates for non-causal explanation rest on or make use of causal information of various sorts. Thus even if it is appropriate to think of these explanations as non-causal, they will often be intertwined with and dependent on causal information.

As an illustration, consider explanations that appeal to facts about the structure of networks in ecology, neurobiology, molecular biology, and other disciplines, as described in Huneman (2010). In many cases such networks are represented by undirected graphs and (I agree) in some cases there is a *prima facie* case for thinking of these as figuring in non-causal explanations. However, when we ask about the evidence which forms the basis for the construction of the networks or what the networks represent, it seems clear they rest on causal information. For example, an undirected network in ecology may represent predator/prey interactions (with the undirected character implying that it does not matter which nodes correspond to predators and which to the prey). Such interactions (on the basis of which the graph is constructed) are certainly causal even if one thinks of the graph itself (perhaps in part because of its undirected character) as providing a non-causal explanation. Similarly, a network model in neurobiology, again represented by an undirected graph, may be constructed on the basis of information about which neural regions causally influence others, so that the network is understood as not merely representing correlational or structural information, although it does not represent causal direction. I do not conclude from this the explanations provided by these models are all causal, but the examples illustrate the extent to which causal and non-causal information can be intertwined in explanatory contexts. This provides another reason for not neglecting causal explanation in our discussion of the non-causal variety.

Before turning to details, two more preliminary remarks: First, my focus will be entirely on possible forms of explanation of empirically contingent claims about the natural world. It may well be, as a number of writers have claimed, that there are mathematical explanations of purely mathematical results—e.g., proofs of such results that are (mathematically) explanatory and which contrast in this respect with other (valid) proofs that are not mathematically explanatory, but I will not address this possibility in this chapter.

Second, the notion of explanation (as captured by the English word and its cognates in many other languages) has, pre-analytically, rather fuzzy boundaries, particularly

when one moves beyond causal explanation. This vagueness encourages the use of what might be described as an “intuitionist” methodology in discussions of non-causal explanation; an example is presented and the reader is in effect asked whether this produces any sense of understanding—an “aha” feeling or something similar. It is not always easy to see what turns on the answer one gives to this question. I have found it difficult to entirely avoid this intuition-based manner of proceeding but in my view it should be treated with skepticism unless accompanied by an account of what is at stake (in terms of connections with the rest of scientific practice or goals of inquiry) in labeling something an explanation. In some cases, as with the explanations of irrelevance considered in section 5, such connections seem obvious enough; in other cases (such as Mother and the strawberries—cf. section 4) not so much.

2. An Interventionist Account of Causation and Causal Explanation

2.1 *Interventions and counterfactual dependence*

According to Woodward (2003), causal claims must correctly describe patterns of counterfactual dependence between variables playing the role of causes and variables playing the role of effects. The relevant notion of counterfactual dependence is understood in terms of *interventions*: *C* causes *E* if and only if there is a possible intervention that changes *C* such that under that intervention, *E* would change. An intervention can be thought of as an idealized experimental manipulation which changes *C* “surgically” in such a way that any change in *E*, should it occur, will occur only “through” the change in *C* and not via some other route. For our purposes, we may think of a *causal explanation* as simply a structure that exhibits or traces such a pattern of dependence, perhaps with the additional qualification that the exhibition in question must satisfy some sort of non-triviality requirement.¹ When an explanation satisfies this condition, Woodward (2003) described it as satisfying a what-if-things-had-been-different requirement (w-requirement) in the sense that it identifies conditions in its explanans such that if those conditions had been different, the explanandum-phenomenon would have been different. (My label for this requirement now seems to me a bit misleading, for reasons given below.) When the variables cited in a candidate explanans meet this requirement there is an obvious sense in which they are “relevant to” or “make a difference to” the explanandum-phenomenon.

Although Woodward (2003) relied heavily on the idea that explanations work by conveying what-if-things-had-been-different information, virtually nothing was said about such questions as how representationally “realistic” a theory or model must be to

¹ Consider the claim that (2.1) the cause of *E* is the cause of *E*. If *E* has a cause (2.1) is true and some intervention on the cause of *E* will be associated with a change in *E*. Most, though, will regard (2.1) as no explanation of *E*, presumably because it is trivial and uninformative (other than implying that *E* has some cause).

convey such information. This has led some readers (e.g., Batterman and Rice 2014) to interpret the w-requirement as a commitment to the idea that only theories that are realistic in the sense of mirroring or being isomorphic (or nearly so) to their target systems can be explanatory. I don't see the interventionist view as committed to anything like this. Instead, what is crucial is (roughly) this: an explanatory model should be such that there is reasoning or inferences licensed by the model that tell one what would happen if interventions and other changes were to occur in the system whose behavior is being explained. This does not require that the model be isomorphic to the target system or even "similar" to it in any ordinary sense, except in the inference-licensing respect just described. To anticipate my discussion in section 5, a minimal model (and inferences performed within such a model) can be used to explain the behavior of real systems via conformity to the w-requirement even if the minimal model is in many respects highly dissimilar (e.g., of different dimensionality) from the systems it explains. The justification for using the minimal model to explain in this way is precisely that one is able to show that various "what-if" results that hold in the minimal model will also hold for the target system.

Turning now to a different subject, the interventionist account requires that for *C* to cause *E*, interventions on *C* must be "possible". Woodward (2003) struggled, not particularly successfully, to characterize the relevant notion of possibility. I will not try to improve on what I said there but will assume that there are some clear cases in which we can recognize that interventions are not (in whatever respect is relevant to characterizing causation) possible. An intervention must involve a physical manipulation that changes the system intervened on and there are cases in which we cannot attach any clear sense to what this might involve. Examples discussed below include interventions that change the dimensionality of physical space and interventions that change a system into a system of a radically different kind—e.g., changing a gas into a ferromagnet. We do possess theories and analyses that purport to tell us how certain systems would behave if they had different spatial dimensions or were a ferromagnet rather than a gas but I assume that such claims should not be interpreted as having to do with the results of possible interventions, but rather must be understood in some other way.

2.2 Invariance

As described above, the characterization of causal explanation does not require that this explicitly cites a *generalization* connecting cause and effect. Nonetheless, in many, perhaps most scientific contexts, generalizations (laws, causal generalizations, etc.), explicitly describing how the explanandum-phenomenon depends on conditions cited in the explanans, are naturally regarded as part of explanations that the various sciences provide. According to Woodward (2003), if these generalizations represent causal relations, they must satisfy *invariance* requirements: for example, at a minimum, such generalizations must be invariant in the sense that they will continue to hold under some range of interventions on factors cited in the explanans. Often, of course, we expect (and find) more in the way of invariance in successful explanations than the

minimal condition described above: we are able to construct explanations employing generalizations which are invariant *both* under a *wide range* of interventions on the variables cited in the explanans, *and* under changes in other variables and conditions not explicitly cited in the explanans—what we may call *background conditions*. Note that, as characterized so far, invariance claims are understood simply as empirical claims about the stability of relationships under variations in the values of various sorts of variables, including variations due to interventions. We will consider below various possibilities for broadening the notion of invariance to include stability under other sorts of variations, including those that do not involve interventions but are rather conceptual or mathematical in character.

2.3 Causal relationships distinguished from conceptual and mathematical relationships

Woodward (2003) (tacitly and without explicit discussion) adopted the common philosophical view that causal (and causal explanatory) relationships contrast with relationships of dependence that hold for purely conceptual, logical, or mathematical reasons.² To employ a standard illustration, Xantippe's widowhood (W) "depends" in some sense on whether Socrates dies (S) but the dependence in question appears to be conceptual (or the result of a convention) rather than causal—one has the sense that (S) and (W) are not distinct in the right way for their relationship to qualify as causal. This is so even though there is an obvious sense in which it is true that by manipulating whether or not Socrates dies, one can alter whether Xantippe is a widow. Thus we should think of the interventionist characterization of causation and causal explanation described above as coming with the rider/restriction that the candidates for cause and effect should not stand in a conceptual or logico-mathematical relationship that is inconsistent with causal interpretation and that, in the case of causal explanation, the explanation should "work" by appealing to a relationship that does not hold for purely conceptual reasons. This contrast between conceptual/mathematical and causal relationships will figure importantly in my discussion below since some plausible candidates for non-causal explanations seem to involve relations between explanans and explanandum that are non-causal because "mathematical" or "conceptual".

2.4 Interventionism is a permissive account of causation

The account of causation and causal explanation described above is broad and permissive—any (non-conceptual) relationship involving intervention-supporting counterfactual dependencies counts as causal, even if it lacks features that other accounts claim are necessary for causation. For example, there is no requirement that causal claims or explanations must provide explicit information about the transfer of energy and momentum or trace processes through time. Similarly, variables that are abstract, generic, multiply realizable, or "upper level" can figure in causal relationships

² For additional discussion of some the subtleties surrounding this notion, see Woodward (2016a).

as long as these variables are possible targets for intervention and figure in intervention-supporting relations of counterfactual dependence. The diagonal length of a square peg can figure in a causal explanation of its failure to fit into a circular hole of a certain diameter (with no reference to the composition of the peg or the forces between its component molecules being required) as long as it is true (as it presumably is) that there are possible interventions that would change the shape of the peg with the result that it fits into the hole.

Summarizing, the picture of causal explanation that emerges from these remarks has the following features: (i) causal explanations provide answers to what-if-things-had-been-different questions by telling us how one variable Y will change under (ii) interventions on one or more others ($X_1, \dots, X_n$). Such interventions must be “possible” in the sense that they correspond to conceptually possible or well-defined physical manipulations. As discussed below, explanations having the structure described in (i) and (ii) will also provide, indirectly, information about what factors do not make a difference to or are irrelevant to the explanandum, but in paradigmatic causal explanations, it is difference-making information that does the bulk of the explanatory work. Finally, (iii) when the relationship between $X_1, \dots, X_n$ and Y is causal, it will be invariant in the sense of continuing to hold (as an empirical matter and not for purely mathematical or conceptual reasons) under some range of interventions on $X_1, \dots, X_n$ and some range of changes in background conditions.

Relaxing or modifying (i)–(iii) either singly or in combination yields various possible candidates for forms of non-causal explanation, which will be explored in subsequent sections. For example, one possible form of non-causal explanation answers w-questions (thus retaining (i)), but does not do so by providing answers to questions about what happens under *interventions*, instead substituting claims about what would happen under different sorts of changes in $X_1, \dots, X_n$ —e.g., changes that correspond to a purely mathematical or conceptual variation not having an interpretation in terms of a possible physical intervention, as in Bokulich (2011) and Rice (2015), among others. Another possible form of non-causal explanation involves retaining (i) and (ii) but dropping requirement (iii), or perhaps retaining (i) but dropping both (ii) and (iii). Here one countenances “explanations” that answer w-questions, but do so by appealing to mathematical, non-empirical relationships. Yet another possibility is that there are forms of explanation that do not tell us anything about the conditions under which the explanandum-phenomenon would have been different, as suggested in Batterman and Rice (2014). (These include the explanations of irrelevance discussed in section 5.)

3. Non-Causal Explanations Not Involving Interventions

Woodward (2003) briefly considered the following candidate for a non-causal explanation. It is possible to show that given assumptions about what the gravitational

potential would be like in an n -dimensional space (in particular, that the potential is given by an n -dimensional generalization of Poisson's equation), Newton's laws of motion, and a certain conception of what the stability of planetary orbits consists in, it follows that no stable planetary orbits are possible for spaces of dimension $n \geq 4$. Obviously orbits of any sort are impossible in a space for which $n = 1$, and it can be argued that $n = 2$ can be ruled out on other grounds, leaving $n = 3$ as the only remaining possibility for stable orbits. Is this an explanation of why stable planetary orbits are possible (in our world)?

Let's assume that this derivation is sound.³ Presumably even if one countenances talk of what would happen under merely possible interventions, the idea of an intervention that would change the dimensionality of space takes us outside the bounds of useful or perhaps even intelligible application of the intervention concept: it is unhelpful, to say the least, to interpret the derivation described above as telling us what would happen to the stability of the planetary orbits under an intervention changing the value of n . Nonetheless one might still attempt to interpret the derivation as answering a *w*-question—it tells us how the possibility of stable orbits (or not) would change as the dimensionality of space changes. In other words, it might be claimed that the derivation satisfies some but not all of the requirements of the interventionist model of causal explanation—it exhibits a pattern of dependence of some kind (perhaps some non-interventionist form of counterfactual dependence) between the possibility of stable orbits and the dimensionality of space, even though this dependence does not have an interventionist interpretation. And since it seems uncontroversial that one of the core elements in many explanations is the exhibition of relationships showing how an explanandum depends on its associated explanans, one might, following a suggestion in Woodward (2003), take this to show that the derivation is explanatory.

Moreover, if it is correct that causal explanations involve dependence relations that have an interventionist interpretation, one might take this to show that the derivation is a case of non-causal explanation—in other words, that one (plausible candidate for a) dividing line between causal and non-causal explanation is that at least some cases of the latter involve dependencies (suitable for answering *w*-questions) that do not have an interventionist interpretation.⁴ Put differently, the idea is that the dependence component in explanation and the interventionist component are separable; drop the latter and retain the former, and you have a non-causal explanation. Suggestions along broadly these lines have been made by a number of writers, including Bokulich (2011),

³ For discussion and some doubts about the soundness claim, see Callender (2005).

⁴ It is worth emphasizing that the candidate explanandum in this case is the *possibility* or not of stable orbits. A natural thought is that if stable orbits are possible, then whether or not some particular planetary orbit is stable is the sort of thing that might be explained causally, but that the possibility of stable orbits is not the sort of thing that can be a causal effect or a target of causal explanation. (The underlying idea would be that causal explanations have to do with what is actual or not, rather than what is possible or impossible.) I lack the space to explore this idea here.

Rice (2015), Saatsi and Pexton (2012), and Reutlinger (2016). For example, Reutlinger argues that explanations of the universal behavior of many very different substances (including gases and ferromagnets) near their critical points in terms of the renormalization group (RG) exhibit the pattern above—the RG analysis shows that the critical point behavior “depends upon” such features of the systems as their dimensionality and the symmetry properties of their Hamiltonians, but the dimensionality of the systems and perhaps also the symmetry properties of their Hamiltonians are not features of these systems that are possible objects of intervention.⁵ In both the case of the stability of the solar system and the explanation of critical point behavior, the “manipulation” that goes on is mathematical or conceptual, rather than possibly physical—e.g., in the former case one imagines or constructs a model in which the dimensionality of the system is different and then calculates the consequences, in this way showing what difference the dimensionality makes. Similarly, in the RG framework, the investigation of the different fixed points of Hamiltonian flows that (arguably) reveal the dependence of critical phenomena on variables like spatial dimensionality does not describe physical transformations of the systems being analyzed, but rather transformations in a more abstract space.

Let us temporarily put aside issues about the structure of the RG explanation (and whether its structure is captured by the above remarks) and focus on the candidate explanation for the stability of the planetary orbits. There is an obvious problem with the analysis offered above. One role that the notion of an intervention plays is that it excludes forms of counterfactual dependence that do not seem explanatory. For example, as is well known, there is a notion of counterfactual dependence (involving so-called backtracking counterfactuals) according to which the joint effects of a common cause counterfactually depend on one another but this dependence is not such that we can appeal to the occurrence of one of these effects to explain the other. In the case of ordinary causal explanation, requiring that the dependence have an interventionist interpretation arguably rules out these non-explanatory forms of counterfactual dependence. The question this raises is whether non-explanatory forms of counterfactual dependence can also be present in candidates for non-causal explanation (thus rendering them non-explanatory) and, if so, how we can recognize and exclude these if we don’t have the notion of an intervention to appeal to.

To sharpen this issue, let me add some information that I have so far suppressed: one may also run the derivation described above backwards, deriving the dimensionality of space from the claim that planetary orbits are stable and assumptions about the gravitational potential and the laws of motion. Indeed, the best-known derivations in the physics literature (such as those due to Ehrenfest 1917 and Buchel 1969) take this second form. Moreover, they are explicitly presented as claims about explanation: that is, as claims that the stability of the planetary orbits *explains* the three-dimensionality

⁵ These claims are not uncontroversial—they are rejected, for example, by Batterman and Rice (2014).

of space.⁶ The obvious question this raises is: which, if either, of these facts (dimensionality, stability) is correctly regarded as the explanans and which as the explanandum? Is it perhaps possible both for stability to explain dimensionality and conversely, so that non-causal explanation can be (sometimes) a symmetric notion? On what basis could one decide these questions?

As Callender (2005) notes, the claim that the stability of the orbits explains the three-dimensionality of space is generally advocated by those with (or at least makes most sense within the context of the assumption of) a commitment to some form of relationalism about spacetime structure: if one is a relationist, it makes sense that facts about the structure of space should “depend” on facts about the possible motions of bodies and the character of the force laws governing those bodies. Conversely, if one is a substantivalist one will think of facts about the structure of space as independent of the motions of bodies in them, so that one will be inclined to think of the direction of explanation in this case as running from the former to the latter.

Without trying to resolve this dispute, let me note that independence assumptions (about what can vary independently of what else) of an apparently non-causal sort seem to play an important role in both purported explanations.⁷ In the case in which the dimensionality of space is claimed to explain the stability of the explanatory orbits, it is assumed that the form of the equation for the gravitational potential is independent of the dimensionality of space in the sense that an equation of the same general form would hold in higher dimensional spaces. Similarly, Newton’s laws of motion are assumed to be independent of the dimensionality of space—it is assumed that they also hold in spaces of different dimensions, with the suggestion being that in such a different dimensioned space ($n \neq 3$), the orbits would not be stable. In the case in which the explanation is claimed to run from the (possible) stability of the orbits to the dimensionality of space, the apparent assumption is that the form of the gravitational potential and the laws of motion are independent of the stability of the orbits in the sense that the former would hold even if the planetary orbits were not possibly stable (in which case the apparent suggestion is that the dimensionality of space would be different). I confess that I find it hard to see what the empirical basis is for either of these sets of claims, although the first strikes me as somehow more natural. As I note below, in other cases of putative non-causal explanations (such as the Königsberg bridge case), there seems to be a more secure basis for claims about explanatory direction.

⁶ Buchel’s paper is entitled, “Why is Space Three-Dimensional?”

⁷ Independence assumptions also play an important role in judgments of *causal* direction—see Woodward (2016b). On this basis one might conjecture that if there is some general way of understanding such assumptions that is not specifically causal, this might be used in a unified theory of causal and non-causal explanation: roughly the idea would be that if X and Y are independent and Z is dependent on X and Y , then the direction of explanation runs from X and Y to Z , and this holds for non-causal forms of (in)dependence.

4. Non-Causal Explanations Involving Mathematical Dependencies but with Manipulable Explanatory Factors

In section 3 we considered putative cases of non-causal explanation in which the explanans factors do not seem to be possible targets for interventions, but in which the relationship between the explanans and explanandum essentially involves assumptions that, however their status is understood, are not a priori mathematical truths. In the example involving the stability of the planetary orbits, the assumption that the gravitational potential for n dimensions takes the form of a generalization of Poisson's equation is not a mathematical truth and similarly for the assumption that the Newtonian laws of motion hold in spaces of dimensionality different from 3. (It is hard to understand these assumptions except as empirical claims, even if it is unclear what empirical evidence might support them.)

I now want to consider some cases that have something like the opposite profile: at least some of the variables figuring in the candidate explanans are possible targets for manipulation (although one might not want to regard the manipulations as interventions in the technical sense, for reasons described in footnote 4) but the *connection* between these and the candidate explanandum seems (in some sense) purely mathematical. Marc Lange has described a simple example which arguably has this structure:

That Mother has three children and twenty-three strawberries, and that twenty-three cannot be divided evenly by three, explains why Mother failed when she tried a moment ago to distribute her strawberries evenly among her children without cutting any. (Lange 2013)

Here a mathematical fact (that 23 cannot be divided evenly by 3) is claimed to explain Mother's failure on this particular occasion. (And, one might think, if this is so, this mathematical fact also explains why Mother always fails on every occasion and why equal division is "impossible".)

Without trying to decide immediately whether this is an "explanation", let's see how it might be fitted into the framework we are using. There is presumably no problem with the notion of manipulating the number of strawberries available to Mother. Perhaps all of the available strawberries must be drawn from a basket and we can add or remove strawberries from the basket. As we vary the number in the basket, we find, e.g., that adding 1 to the 23 makes even division possible, subtracting 1 makes it impossible and so on. The overall pattern that emerges is that even division among the children is possible when and only when the number of strawberries is evenly divisible by three. It is not a huge stretch to think that the observation that this pattern holds fits naturally (in this respect) into the *w*-question framework and that the observation thus isolates a factor on which the explanandum (whether even division is possible) "depends". On these grounds one might think that an interesting similarity is present between this example and more paradigmatic cases of causal

explanation and that this warrants regarding the example as providing a genuine explanation.⁸

Of course there is also the obvious disanalogy mentioned earlier: given the particular facts in the example (number of strawberries and children) the connection between these and the candidate explanandum (whether equal division is possible) follows just as a matter of mathematics, without the need for any additional assumptions of a non-mathematical nature. Presumably this is why it does not seem correct to think of the relationship between the particular facts cited in the candidate explanans and the failure to divide, or impossibility of dividing equally as causal. Instead, as in the case of the relationship between Socrates' death and Xantippe's widowhood, it seems more natural to express the dependence between the possibility of equal division and the number of strawberries and children by means of locutions like "brings about by" that are appropriate for cases of non-causal dependence: by varying the number of strawberries or children one brings it about that Mother succeeds or fails at equal division.

Our reaction to this example may be colored by the fact that the mathematical fact to which it appeals is trivial and well known; this may contribute to the sense that many may have that in this case citing the mathematical fact does not greatly enhance understanding, so that (at best) only in a very attenuated sense has an explanation been provided. However, there are other cases, such as the well-known Königsberg bridge problem, which seem to have a similar structure where many will have more of a sense that an explanation has been furnished. Suppose we represent the configuration of bridges and land masses in Königsberg by means of an undirected graph in which bridges correspond to edges, and the land masses they connect to nodes or vertices. An *Eulerian path* through the graph is a path that traverses each edge exactly once. Euler proved that a necessary condition for a graph to contain an Eulerian path is that the graph be connected (there is a path between every pair of vertices) and that it contain either zero or two nodes of odd degree, where the degree of a node is the number of edges connected to the node.⁹ This condition is also sufficient for a graph to contain an Eulerian path. The Königsberg bridge configuration does not meet this condition—each of the four land masses is connected to an odd number of bridges—and it follows that it contains no Eulerian path.

One might think of this demonstration in the following way: we have certain contingent facts—the connection pattern of the bridges and land masses of Königsberg. Given these, one can derive via a mathematical argument that makes use of no additional

⁸ For a similar treatment of this example, see Jansson and Saatsi (forthcoming).

⁹ This is unmythical when you think about it. Except for the starting and end point of the walk, to traverse an Eulerian path one must both enter each land mass via a bridge and exit via a different bridge. If each bridge is to be traversed exactly once, this requires that each such non-terminal land mass must have an even number of edges connected to it. At most two land masses can serve as starting and end points, with an odd number of edges connected to them. It is interesting to note (or so it seems to me) that it is a *proof* or argument along lines like this which does whatever explanatory work is present in the example rather than just the specification of the difference-making conditions itself.

empirical premises that it is impossible to cross each bridge exactly once. (That is, the connection between explanans and explanandum is entirely mathematical rather than empirical.) Moreover, the derivation makes use of information that can be used to answer a number of *w*-questions about the explanandum—as just one sort of possibility, the derivation tells us about alternative possible patterns of connectivity which would make it possible to traverse an Eulerian path among the bridges as well as about other patterns besides the actual one in which this would not be possible. In doing this the explanation also provides information about the many features of the situation that do not matter for (are irrelevant to) whether it is possible to traverse each bridge exactly once: it does not matter where one starts, what material the bridges are made of, or even (as several writers note) what physical laws govern the bridges, as long as they provide stable connections. These assertions about the irrelevance of physical detail are bound up with our sense that Euler's analysis isolates the abstract, graph-theoretical features of the situation that are relevant to whether it is possible to traverse an Eulerian path. Note, however, that this information about irrelevance figures in the analysis only against the background of information about what *is* relevant, which has to do with the connectivity of the graph.

Note also that despite this mathematical connection between explanans and explanandum, the notion of changing or manipulating the bridge configuration—e.g., by constructing additional bridges or removing some—and tracing the results of this does not seem strained or unclear. This also fits naturally with an account of the example in terms of which it is explanatory in virtue of providing information to *w*-questions.

It is also worth noting that in this case, in contrast to the example involving the dimensionality of space in section 3, the direction of the dependency relation seems unproblematic. The configuration of the bridges has perfectly ordinary causes rooted in human decisions to construct one or another particular configuration. Because these decisions cause the configuration, it is clear that the impossibility of traversing an Eulerian path is not somehow part of an explanation of the configuration. Rather, if this is a case of explanation, the direction must run from the configuration to the impossibility of traversing, with the configuration instead having the causes described above. This shows one way in which the problem of distinguishing explanatory from non-explanatory patterns of dependence in connection with candidates for non-causal explanation might be addressed.

5. The Role of Information about Irrelevance in Explanation

As noted above, the *w*-question conception focuses on the role of factors that are explanatorily relevant to an explanandum—relevant in the sense that variations in those factors make a difference to whether the explanandum holds. A number of recent discussions have instead focused on what might be described as the role of

irrelevance or independence in explanation—on information to the effect that some factors do *not* make a difference to some explanandum, with some writers seeming to suggest that some explanations work primarily or entirely by citing such independence information and that interventionist and other difference-making accounts of explanation cannot accommodate this fact (see, e.g., Batterman and Rice 2014; Gross 2015). Indeed, it sometimes seems to be suggested that some explananda can be explained by citing *only* factors that are irrelevant to it, with difference-making factors playing no role at all. In this section I want to explore some issues raised by the role of information about irrelevance.

First we need to clarify what is meant by the notions of relevance and irrelevance. The interventionist understanding of relevance is that X is relevant to Y as long as *some* interventions that change the value of Y are associated with changes in X ; X is irrelevant to Y if there are no such interventions. Suppose that X and Y can take a range of different values and that X and Y are related by $F(X) = Y$. Assume F specifies that some changes in X are associated with changes in Y and that others are not—in other words, F is not a 1–1 function, although it does not map all values of X into the same value of Y . In such a case, X is relevant to Y , although of course we may also go on to describe more specifically which changes in the value of X are relevant to Y and which others are not. My understanding of the what-if-things-had-been-different idea has always been that in such cases F provides w-information and is explanatory in virtue of describing the pattern of dependence of Y on X even though that pattern is such that some changes in X make no difference to the value of Y .¹⁰ We may also generalize the w-account to include (in) dependence information that is not understood in terms of interventions, as suggested above, in which case similar remarks (e.g., that the dependence need not be 1–1) apply.

As suggested in passing above, information about independence and irrelevance is in many ways the flip side of the dependence or relevance information emphasized in the interventionist account, since the latter can often be “read” off from the former. To take the most obvious possibility, when (5.1) some variable Y is represented as dependent on others $X_1, \dots, X_n$, this (at least often) implicitly conveys that other variables $Z_1, \dots, Z_n$, distinct from $X_1, \dots, X_n$ that are not explicitly mentioned in (5.1) are irrelevant to Y . When the gravitational inverse square law and Newton’s laws of motion are used to explain the trajectories of the planets, one does not have to explicitly add the information that the colors of the planets are irrelevant to their trajectories since the use of these laws conveys this information. In these respects, virtually all

¹⁰ I mention this because some writers (e.g. Gross 2015) interpret me as holding the contrary view that when the relation between X and Y is not 1–1, this relationship is not explanatory (because it is not a dependence or difference-making relationship). Gross describes a biological example in which (put abstractly) some changes in the value of X are relevant to Y and many others are not; he claims that in this case the interventionist account cannot capture or take notice of the biological significance of this irrelevance information. My contrary view is that this is an ordinary dependence or difference-making relation and, according to interventionism, explanation can proceed by citing this relationship.

explanations convey information about irrelevance—this is not a distinctive feature of some specific subclass of explanations.

In examples like the one just described there is an obvious sense in which the dependence information seems to be doing the explanatory work, with the independence information following derivatively from the dependence information. One indication of this is that the independence information by itself, apart from dependence information, does not seem explanatory: Presumably no one would be tempted to think that one could explain the motions of the planets just by citing information to the effect that factors such as color are irrelevant to that motion.

In other cases, however, one has the sense that information about independence or irrelevance may be playing a more substantial explanatory role. Consider an *equilibrium explanation*, where one component of the explanation involves showing that the outcome being explained would result from a large number of possible initial states. As an illustration, suppose the final state of a gas (e.g., that it exerts a certain equilibrium pressure after being allowed to diffuse isothermally into a fixed volume) is explained by means of a demonstration that almost all initial states of the gas compatible with certain macroscopic thermodynamic constraints (e.g., the temperature of the gas and the volume of the container) will evolve to the same equilibrium outcome. Another illustration is provided by Fisher's well-known explanation of sex allocation among offspring and the various generalizations of this due to Hamilton, Charnov, and others, where the factors influencing equilibrium outcomes are shown to be independent of the details of specific episodes of fertilization.

Such explanations are often claimed to be non-causal or not captured within a difference-making framework since they do not involve tracing the actual trajectory of the specific events leading to the explanandum-outcome. In a brief discussion, Woodward (2003) objected to the characterization of such explanations as non-causal: virtually always such explanations do invoke dependency or difference-making information (which can be understood in terms of interventionist counterfactuals) in addition to information to the effect that many initial states will lead to the same outcome. For example, in the case of the gas, the pressure will of course depend on the maintained temperature and the container volume—vary these and the temperature will vary. Since these dependency relations can be given an interventionist interpretation, one can interpret the explanations as providing causal explanations of why one particular equilibrium rather than another obtains.

While these observations still seem to me to be correct, I think it is also true that independence or irrelevance information seems to play a somewhat different role in these equilibrium explanations than it does in the explanation of planetary trajectories, where its role seems essentially trivial. Perhaps one aspect of the difference (not the only difference as discussed below) is this: a property like color plays no interesting role anywhere in mechanics or in most of the rest of physics and no one will be surprised by the observation that the influence of gravity on the planets is unaffected by their color. Indeed, the question: why does color not matter for planetary trajectories? does not

seem to arise in any natural way nor is it obvious what would serve as an answer to it. On the other hand, facts about the detailed trajectories of individual molecules are among the sorts of facts that physics pays attention to: they are relevant to what happens in many contexts and are explananda for many physical explanations. There thus seems to be a live question about why, to a very large extent, details about individual molecular trajectories don't matter for the purposes of predicting or explaining thermodynamic variables. Replacing details about the individual trajectories of the 10^{23} molecules making up a sample of gas with a few thermodynamic variables involves replacing a huge number of degrees of freedom with a very small number which nonetheless are adequate for many predictive and explanatory purposes. It is natural to wonder why this "variable reduction" strategy works as well as it does and why it is that, given the values of the thermodynamic variables, further variations in the molecular trajectories almost always make no difference to many of the outcomes specifiable in terms of thermodynamic variables.

Here we seem to be asking a different kind of question than the questions about the identification of difference-makers that characterize straightforward causal analysis; we are asking instead why variations in certain factors do *not* make a difference to various features of a system's behavior, at least given the values of other factors. Put slightly differently, we are still interested in w-questions but now our focus is on the fact that if various factors had been different in various ways, the explanandum would *not* have been different and perhaps on understanding why this is the case.¹¹ (I have so far not tried to provide any account of what such an explanation would look like—that will come later.)

Note, however, that these observations do *not* support the idea that one can explain why some outcome occurs by just citing factors that are irrelevant to it. In the example above and others discussed below, it seems more natural to regard the claims about irrelevance as explananda (or at least as claims that are in need of justification on the basis of other premises) rather than as part of an explanans (or premises that themselves do the explaining or justifying). That is, rather than citing the irrelevance of V to E in order to explain E , it looks as though what we are interested in explaining or understanding is why V is irrelevant to E . Explaining why V is irrelevant to E is different from citing the irrelevance of V to explain E . Moreover, independently of this point, in the examples we have been looking at, the irrelevance of certain factors to some outcome is conditional on the values of other factors that *are* identified as relevant, with the form of the explanatory claim being something like this: (5.1) Given the values of variables $X_1, \dots, X_n$ (which are relevant to outcome E)—e.g., temperature and volume—variations in the values of additional variables $V_1, \dots, V_n$ (e.g., more detailed facts about individual

¹¹ This is why I said earlier that my use of the phrase "w-information" in Woodward (2003) was a bit misleading or imprecise: I had in mind the specification of changes in factors in an explanans under which the explanandum would have been different but of course it may be true that under some changes in the explanans factors, the explanandum would not have been different.

molecular trajectories) are irrelevant to E .¹² Thus insofar as the irrelevant variables or the information that they are irrelevant have explanatory import, they do so in the context of an explanation in which other variables *are* relevant.

What might be involved in explaining that certain variables are irrelevant to others (or irrelevant to others conditional on the values of some third set of variables)? Although several writers, including Batterman and Rice (2014), defend the importance of such explanations and offer examples, I am not aware of any fully systematic treatment. Without attempting this, I speculate that one important consideration in many such cases is that there is an underlying dynamics which, even if it is not known in detail, supports the claims of irrelevance—what we want is insight into how the working of the dynamics makes for the irrelevance of certain variables. For example, in Fisher's well-known treatment of sex allocation, it is not just that many fertilization episodes that differ in detail can be realizers of the creation of females or males.¹³ The equilibria in such analyses are (or are claimed to be) *stable* equilibria in the sense that perturbations that take populations away from equilibrium allocations are soon returned to the equilibrium allocation because of the operation of natural selection—it being selectively disadvantageous to produce non-equilibrium sex ratios. In other words, there is a story to be told about the structure of the dynamics, basins of attraction, flows to fixed points, etc. that gives us insight into why the details of individual episodes do not matter to the outcome. Similarly for the behavior of the gas. There is nothing similar to this in the case of explaining the irrelevance of colors to the trajectories of planets, which is why it is hard to see what non-trivial form such an explanation would take.

In the cases considered so far in this section the notion of irrelevance has an obvious interventionist interpretation. However, there are other cases, discussed below, in which we need to broaden the notions of relevance and irrelevance to include reference to variations or changes that do not have an interventionist interpretation or where it is at least not obvious that such an interpretation is appropriate. These include cases in which it follows as a matter of mathematics that, given certain generic constraints, variations in values of other variables or variations in structural relationships make no difference to some outcome, but where the variations in question are not (or may not be) the sort of thing that can be produced by interventions.

A possible illustration is provided by the use of the method of arbitrary functions and similar arguments to explain the behavior of gambling devices such as roulette wheels. An obvious explanatory puzzle raised by such devices is to understand why they produce stable frequencies of outcomes strictly between 0 and 1 despite being deterministic, and despite the fact that the initial conditions characterizing any one device will vary from trial to trial (and of course also vary across devices) and that

¹² For further discussion of this sort of conditional irrelevance (as I call it) see Woodward (forthcoming).

¹³ This is one reason (of several) why thinking of such examples (just) in terms of multiple realizability misses important features.

different devices are governed by different detailed dynamics. Moreover, these relative frequencies are also stable in the sense that they are unaffected by the manipulations available to macroscopic agents like croupiers. Very roughly, it can be shown that provided that the distribution of initial conditions on successive operations of such devices satisfies some generic constraints (e.g., one such constraint is that the distribution is absolutely continuous) and the dynamics of the devices also satisfy generic constraints, the devices will produce (in the limit) outcomes with well-defined probability distributions and stable relative frequencies—in many cases (when appropriate symmetries are satisfied) uniform distributions over those outcomes. It is natural to think of these sorts of analyses as providing explanations of the facts about irrelevance and independence described above—why the manipulations of the croupier do not matter to the distribution of outcomes and so on.

In such cases it is not clear that all of the variations under which these devices can be shown to exhibit stable behavior have an interventionist interpretation. For example, the information that any one of a large range of different dynamics would have generated the same behavior seems to have to do with the consequences of variations within a mathematical space of possible dynamics rather than with variations that necessarily have an interventionist interpretation. Relatedly, it is arguable that those features of the system that the analysis reveals as relevant to the achievement of stable outcomes—the generic constraints on the initial conditions and on the dynamics—are not naturally regarded as “causes” of that stability in the interventionist sense of cause. For example, it is not obvious that the fact that the distribution of initial conditions satisfied by some device is absolutely continuous should count as a “cause” of the device’s behavior. On the other hand, if we follow the line of thought in previous sections and extend the notion of information that answers *w*-questions to include cases in which the information in question does not have to do with interventionist counterfactuals but rather with what happens under variations of different sorts (in initial conditions, dynamics, etc.) and where the answer may be that some outcome or relationship does *not* change under such variation (i.e., the variations are irrelevant) we can accommodate examples of this sort. That is, we can think of these as explanations of irrelevance where the irrelevance in question is irrelevance under variations of a certain sort but where the variations do not have an interventionist interpretation. In such cases, irrelevance is demonstrated mathematically by showing that the mathematical relationships between the variations and some phenomenon or relationship is such that the latter does not change under the former.

I conclude this section by briefly exploring some additional issues about irrelevance in the context of some recent claims made by Batterman and Rice (2014) about minimal models and their role in explanation. Abstractly speaking, we can think of a minimal model as a model which captures aspects of the common behavior of a class of systems (and of the behavior of more detailed models of such systems in this class). A minimal model serves as a kind of stand-in for all of the systems for which it is a minimal model—for an appropriate class, results that can be shown to obtain for the

minimal model must also hold for other models and systems within the delimited class, no matter what other features they possess. Thus one can make inferences (including “what if” inferences) and do calculations using the minimal model, knowing that the results “must” transfer to the other models and systems. Here the “must” is mathematical; one shows as a matter of mathematics that the minimal model has the stand-in or surrogate role just described with respect to the other models and systems in the universality class. Renormalization group analysis (RGA) is one way of doing this—of justifying the use of a minimal model as a surrogate. In this respect, RGA delimits the “universality class” to which the minimal model belongs.

A striking example, discussed by Batterman and Rice, is provided by a brief paper by Goldenfeld and Kadanoff (1999) which describes the use of a minimal model for fluid flow (the lattice gas automaton or LGA). The model consists of point particles on a two-dimensional hexagonal lattice. Each particle interacts with its nearest neighbors in accord with a simple rule. When this rule is applied iteratively and coarse-grained averages are taken, a number of the macroscopic behaviors of fluids are reproduced. As Goldenfeld and Kadanoff explain, the equations governing macroscopic fluid behavior result from a few generic assumptions: these include locality (the particles making up the fluid are influenced only by their immediate neighbors), conservation (of particle number and momentum), and various symmetry conditions (isotropy and rotational invariance of the fluid). These features are also represented in the LGA and account for its success in reproducing actual fluid behavior, despite the fact that real fluids are not two-dimensional, not lattices and so on.

Batterman and Rice make a number of claims about the use of minimal models in explanation. First, they seem to suggest in one passage that such models are explanatory because they provide information that various details are irrelevant to the behavior of the systems modeled.¹⁴

[The] models are explanatory because of a story about why a class of systems will all display the same large-scale behavior because the details that distinguish them are irrelevant.

(2014: 349)

Elsewhere they write, in connection with the use of the renormalization group to explain critical point behavior:

The fact that the different fluids all possess these common features (having to do with behavior near their critical points) is also something that requires explanation. The explanation of this fact is provided by the renormalization group-like story that delimits the universality class by demonstrating that the details that genuinely distinguish the fluids from one another are irrelevant for the explanandum of interest. (2014: 374)

¹⁴ Batterman informs me that this is not what the quoted passage was intended to express: his idea was rather that what justifies the use of the minimal model for explanatory purposes is the RG story about irrelevance of other actual details omitted by the minimal model.

Second, they claim that the features that characterize the minimal model are not causes of (and do not figure in any kind of causal explanation of) the fluid phenomena being explained:

We think it stretches the imagination to think of locality, conservation, and symmetry as causal factors that make a difference to the occurrence of certain patterns of fluid flow. (2014: 360)

Although this may not be their intention, the first set of passages makes it sound as though they are claiming that the common behavior of the fluids can be explained just by citing factors that are irrelevant to that behavior or by the information that these factors are irrelevant. Let me suggest a friendly amendment: it would be perspicuous to distinguish the following questions: First, (a) why is it justifiable to use this particular model (LGA) as a minimal model for a whole class of systems? Second, (b) why do systems in this class exhibit the various common behaviors that they do? I agree with what I take to be Batterman and Rice's view that the answer to (a) is provided by renormalization-type arguments or more generally by a mathematical demonstration of some kind that relates the models in this class to one another and shows that for some relevant class of behaviors, any model in the class will exhibit the same behavior as the minimal model. I also agree with Batterman and Rice that in answering this question one is providing a kind of explanation of (or at least insight into) why the details that distinguish the systems are irrelevant to their common behavior. But, to repeat an observation made earlier, the explanandum in this case is a claim about irrelevance (what is explained is why certain details are irrelevant); this answer to (a) does not support the contention that irrelevance claims by themselves are enough to explain (b).

Instead, it seems to me that the explanation for why (b) holds is provided by the minimal model itself in conjunction with information along the lines of (a) supporting the use of the minimal model as an adequate surrogate for the various systems in the universality class. Of course the minimal model does not just consist in claims to the effect that various factors are irrelevant to the common behavior of the systems (although its use certainly implies this), so we should not think of this explanation of (b) as consisting just in the citing of irrelevance information. Instead the minimal model also provides information about a common abstract structure shared by all of the systems in the universality class—structure that (as I see it) is relevant to the behavior of these systems. Here, as in previous cases, relevance and irrelevance information work together, with the irrelevance information telling us, roughly, why it is justifiable to use a certain minimal model and why various details that we might have expected to make a difference to systems in the universality class do not and the relevance information identifying the shared structure that does matter.

Regarding this shared structure several further questions arise. First, does the structure furnish a *causal* explanation of (b)? Here I agree with Batterman and Rice that the answer is “no”, or at least that it is “no” given an interventionist account of causation. The features characterizing the structure are just not the sort of things that are well-defined objects of intervention—one cannot in the relevant sense intervene

to make the interactions governing the system local or non-local, to change the dimensionality of the system, and so on. However, I would contend that we should not necessarily infer from this that the minimal model does not cite difference-making factors at all or that these difference-making factors have no explanatory significance; instead it may be appropriate to think of the model as citing non-causal difference-making factors which have explanatory import in the manner that some of the putative explanations in section 3 do. One reason for thinking that something like this must be the case is that the LGA and associated RG-type analyses are not just used to provide insight into why various details distinguishing the systems are irrelevant to certain aspects of their behavior; they are also used to calculate (and presumably explain) various other more specific features of the systems in question—critical exponents, relations among critical exponents, deviations from behavior predicted by other (e.g., mean field) models, and so on. These are not explananda that can be derived or explained just by citing information to the effect that various details are irrelevant or non-difference-makers; one also needs to identify which features are relevant to these behaviors and it is hard to see how this could fail to involve difference-making information, albeit of a non-causal sort.

I thus find plausible Reutlinger's recent suggestion (2016) that explanations of the RG sort under discussion work in part by citing what-if-things-had-been-different information of a non-causal sort. I will add, however, that, for reasons described above, I do not think that this captures the whole story about the structure of such explanations; Batterman and Rice are correct that explanations of irrelevance also play a central role in such explanations.

6. Conclusion

In this chapter I have tried to show how the interventionist account of causal explanation might be extended to capture various candidates for non-causal explanation. These include cases in which there is empirical dependence between explanans and explanandum which does not have an interventionist interpretation, and cases in which the relation between explanans and explanandum is conceptual or mathematical. Examples in which claims about the irrelevance of certain features to a system's behavior are explained or justified are also acknowledged and discussed, but it is contended that difference-making considerations also play a role in such examples.

Acknowledgments

Many thanks to Bob Batterman, Collin Rice, and the editors for helpful comments on earlier drafts.

References

- Batterman, R. and Rice, C. (2014), 'Minimal Model Explanations', *Philosophy of Science* 81: 349–76.
- Bokulich, A. (2011), 'How Scientific Models Can Explain', *Synthese* 180: 33–45.
- Buchel, W. (1969), 'Why is Space Three-Dimensional?', trans. Ira M. Freeman, *American Journal of Physics* 37: 1222–4.
- Callender, C. (2005), 'Answers in Search of a Question: "Proofs" of the Tri-Dimensionality of Space', *Studies in History and Philosophy of Modern Physics* 36: 113–36.
- Ehrenfest, P. (1917), 'In What Way Does It Become Manifest in the Fundamental Laws of Physics that Space Has Three Dimensions?', *Proceedings of the Amsterdam Academy* 20: 200–9.
- Goldenfeld, N. and Kadanoff, L. (1999), 'Simple Lessons from Complexity', *Science* 284: 87–9.
- Gross, F. (2015), 'The Relevance of Irrelevance: Explanation in Systems Biology', in P.-A. Braillard and C. Malaterre (eds.), *Explanation in Biology: An Enquiry into the Diversity of Explanatory Patterns in the Life Sciences* (Dordrecht: Springer), 175–98.
- Huneman, P. (2010), 'Topological Explanation and Robustness in Biological Systems', *Synthese* 177: 213–45.
- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Jansson, L. and Saatsi, J. (forthcoming), 'Explanatory Abstractions', *British Journal for the Philosophy of Science*.
- Reutlinger, A. (2016), 'Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation', *Philosophy of Science* 83: 733–45.
- Rice, C. (2015), 'Moving beyond Causes: Optimality Models and Scientific Explanation', *Noûs* 49: 589–615.
- Saatsi, J. and Pexton, M. (2012), 'Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Non-Causal Explanations', *Philosophy of Science* 80: 613–24.
- Woodward, J. (2003), *Making Things Happen* (New York: Oxford University Press).
- Woodward, J. (2016a), 'The Problem of Variable Choice', *Synthese* 193: 1047–72.
- Woodward, J. (2016b), 'Causation in Science', in P. Humphreys (ed.), *The Oxford Handbook of Philosophy of Science* (New York: Oxford University Press), 163–84.
- Woodward, J. (forthcoming), 'Explanatory Autonomy: The Role of Proportionality, Stability and Conditional Irrelevance'.

PART II

Case Studies from the Sciences

Searching for Non-Causal Explanations in a Sea of Causes

Alisa Bokulich

To anyone who, for the first time, sees a great stretch of sandy shore covered with innumerable ridges and furrows, as if combed with a giant comb, a dozen questions must immediately present themselves. How do these ripples form?

Hertha Ayrton ([1904] 1910: 285)¹

1. Introduction

According to a position we might label causal imperialism, all scientific explanations are causal explanations—to explain a phenomenon is just to cite the causes of that phenomenon.² Defenders of non-causal explanation have traditionally challenged this imperialism by trying to find an example of an explanation for a phenomenon for which no causal explanation is available.³ If the imperialist can, in turn, find a causal explanation of that phenomenon, then it is believed that the defender of non-causal explanation has been defeated.⁴ Implicit in such a dialectic are the following two assumptions: first, that finding an example of a non-causal explanation requires finding something like an uncaused event, and, second, that causal and non-causal explanations of a phenomenon are incompatible. This has left non-causal explanations as relatively few and far between, relegating them to fields such as fundamental physics or mathematics.

¹ This quotation is taken from the first paper ever permitted to be read by a woman at a meeting of the Royal Society of London.

² An example of a defender of such a position is David Lewis (1986), but more often it is a position that is assumed as a default, rather than being explicitly defended. Brad Skow (2014) similarly argues, “what I say here does not prove that there are no possible examples of non-causal explanations, but it does, I think, strengthen the case” (446).

³ This is arguably why defenders of non-causal explanation have primarily looked to examples in mathematics and quantum mechanics, where causal explanations are thought to be excluded.

⁴ As Marc Lange (2013: 498–9) notes, for example in the case of the prime life cycle of cicadas, there is often a causal explanation in the close vicinity of a non-causal explanation that can be conflated if the explananda are not carefully distinguished.

In what follows, I challenge these two assumptions. Non-causal explanations do not require finding a phenomenon for which no causal story can be told. I argue instead that one can have a non-causal explanation of a phenomenon even in cases where a complete causal account of the phenomenon is available. Having a causal explanation of a phenomenon does not preclude also having an alternative, non-causal explanation for that same phenomenon. Causal and non-causal explanations are complementary, and each can be useful for bringing out different sorts of insights.

I begin by introducing my approach to scientific explanation, which includes what I call the “eikonic” alternative to the ontic conception of explanation, and which distinguishes two types of explanatory pluralism. I will then lay out my framework for model-based explanation, within which both causal and non-causal explanations can be understood, and illustrate this framework by very briefly reviewing my previous work on non-causal model explanations. I will then turn to an examination of various proposals in the philosophical literature for what is required for an explanation to count as non-causal. After noting the strengths and weaknesses of these proposals, I will extract what I take to be a core conception of non-causal explanation. I will use as a detailed case study the example of how Earth scientists are explaining the formation of regularly-spaced sand ripples in the subfield known as aeolian geomorphology. I will conclude that even when it comes to familiar, everyday “medium-sized dry goods” such as sand ripples, where there is clearly a complete causal story to be told, one can find examples of non-causal scientific explanations.

2. Model-Based Explanations

Those who defend the causal approach to scientific explanation have traditionally also subscribed—either implicitly or explicitly—to the ontic conception of explanation (e.g., Salmon 1984, 1989; Craver 2007; 2014; Strevens 2008).⁵ According to the ontic conception, explanations just are the full-bodied entities and processes in the world themselves. The claim is that the particular baseball, the particular adrenaline molecules, and the particular photons are not just causes or causally relevant, but that they are further scientific *explanations*. As Carl Craver defines it:

Conceived ontically... the term explanation refers to an objective portion of the causal structure of the world, to the set of factors that produce, underlie, or are otherwise responsible for a phenomenon. Ontic explanations are not texts; they are full-bodied things. They are not true or false. They are not more or less abstract. They are not more or less complete. They consist in all and only the relevant features of the mechanisms in question. There is no question of ontic explanations being “right” or “wrong,” or “good” or “bad.” They just are. (Craver 2014: 40)

In another paper (Bokulich 2016), I have argued that the ontic conception of explanation is highly problematic, if not incoherent. Insofar as one is interested in normative

⁵ It is important to distinguish a *conception* of explanation, which is a claim about what explanations are, from an *account* of explanation, which is a claim about how explanations work.

constraints on scientific explanation, one must reject the ontic conception and instead view scientific explanation as a human activity involving representations of the world.

Elsewhere I have defended a version of the representational view that I call the eikonic conception of explanation, named from the Greek word ‘eikon’ meaning representation or image (Bokulich forthcoming). Like the ontic conception, the eikonic conception is a claim about what explanations are, and is compatible with many different accounts about how explanations work (e.g., causal, mechanistic, nomological, and of course non-causal accounts of explanation). On the eikonic view, a causal explanation involves citing a particular representation of the causal entities, rather than the brute existence of the causal entities themselves. Rejecting the view that explanations just are the causal entities and processes in the world themselves makes room for the possibility of a non-causal explanation even in cases where there is a complete causal story to be had about the production of the phenomenon. As we will see in section 4, a non-causal explanation is an explanation where the explanatory factors cited, the “explanans”, are not a direct representation of the causal entities and processes. This very abstract characterization of a non-causal explanation allows for the possibility of different kinds of non-causal explanation, and will be fleshed out in the context of the case study below.

As suggested by the preceding, a second component of my approach to scientific explanation is a commitment to explanatory pluralism. The expression ‘explanatory pluralism’ has been used to express two different views in the philosophy of science. Originally it was used in opposition to those who argued that all cases of explanation can be subsumed under a single, unitary account, such as the covering-law model or, more recently, the causal account of explanation. Explanatory pluralism in this sense (what I call “type I” explanatory pluralism) is the view that scientists use different types of explanations (at different times or in different fields) with respect to *different* phenomena (e.g., while evolutionary biologists might use the unificationist account of explanation for their explananda, molecular biologists use mechanistic explanations for theirs). More recently, however, explanatory pluralism has come to mean that there can be more than one scientifically acceptable explanation of a single, given phenomenon (what I call “type II” explanatory pluralism). So for example, there could be two explanations for the morphology of a particular river—one that was deductive-nomological in form, while another was mechanistic. Both are scientifically acceptable explanations for why a river has the shape that it does, but they take different forms and appeal to different explanatory factors. Type II explanatory pluralism opens up the possibility that we can have multiple scientific explanations for a phenomenon, some of which are “deeper” than others (e.g., Hitchcock and Woodward 2003). While type I explanatory pluralism has become widely accepted (except perhaps by the causal imperialists), type II explanatory pluralism is more controversial. Type II pluralism not only presupposes type I (that there are different forms of scientific explanation), but goes further in asserting that these different kinds of explanation can be applied

to the *same* phenomenon. I suspect that part of the resistance to type II explanatory pluralism comes from a subtle conflation between ‘cause’ and ‘explain’ that is endemic to the ontic conception. The sense of explanatory pluralism that I will be most concerned with here is type II, insofar as I will be arguing that there can be causal and non-causal explanations for one and the same phenomenon.

A third component of my approach to scientific explanation is my view that many explanations in science proceed by way of an idealized model, in terms of what I have called model-based explanation (Bokulich 2008a, 2008b, 2011). As we will see, both the causal and non-causal explanations of sand ripples, discussed in section 4, are examples of model-based explanation. My account of model-based explanation can be understood as consisting of the following four components. First, the explanans makes central use of a model that (like all models) involves some degree of idealization, abstraction, or even fictionalization of the target. Second, the model explains the explanandum phenomenon by showing how the elements of the model correctly capture the patterns of counterfactual dependence in the target system, allowing one to answer a wide range of what James Woodward (2003) calls “what-if-things-had-been-different” questions (w-questions). Third, there must be a justificatory step by which the model representation is credentialed (for a given context of application) as giving genuine physical insight into the phenomenon being explained; that is, there are good evidential grounds for believing the model is licensing correct inferences in the appropriate way. Explanation is a success term and requires more than just an “Aha!” feeling. Finally, this approach allows for different types of model explanations (e.g., causal, mechanistic, nomic, or structural model explanations) depending on the particular origin or ground of the counterfactual dependence (Bokulich 2008a: 150).

In my previous work on explanations in semiclassical physics, I identified a particular kind of non-causal model explanation that I called structural model explanations (Bokulich 2008a). These particular structural model explanations in semiclassical mechanics involve an appeal to classical trajectories and their stability exponents in explaining a quantum phenomenon known as wavefunction scarring. Wavefunction scarring is an anomalous enhancement of quantum eigenstate intensity along what would be the unstable periodic orbits of a classically chaotic system. Although scarring is a *quantum* phenomenon, the received scientific explanation appeals to the *classical* orbits to explain the behavior of the wavepackets, and the classical Lyapunov exponent to explain the intensity of the scar. According to quantum mechanics, however, there are no such things as classical trajectories or their stability exponents—they are fictions. Insofar as classical periodic orbits do not exist in quantum systems, they cannot enter into causal relations. Hence the semiclassical model explanations that appeal to these trajectories are a form of *non-causal* explanation. In accordance with my generalized Woodwardian approach to model explanation, these semiclassical models are able to correctly capture the patterns of counterfactual dependence in the target system, and the theory of semiclassical mechanics provides the justificatory step, credentialing

the use of these classical structures as giving genuine physical insight into these quantum systems.⁶

Although many might be willing to admit the possibility of non-causal explanations in quantum mechanics, a theory famously unfriendly to causality, the idea that there could be non-causal explanations outside of fundamental physics or mathematics is met with more skepticism. Before arguing that one can find non-causal explanations of familiar macroscopic phenomena like sand ripples, it is important to first clarify what is required for an explanation to count as genuinely non-causal. In section 3, I will show how a core conception of non-causal explanation can be distilled from the recent literature on this topic.

3. What Makes an Explanation Non-Causal?

The generalized Woodwardian approach that I used as a framework capable of encompassing both causal and non-causal explanations has more recently been adopted and further developed in different ways by several scholars defending non-causal explanation, such as Juha Saatsi and Mark Pexton (2013), Collin Rice (2015), Saatsi (2016), and Alexander Reutlinger (2016). Even within this general framework, however, the question still remains what distinguishes specifically *non-causal* explanations. Non-causal explanations are typically defined negatively—as conveying explanatory information in ways other than by citing the causes of the explanandum phenomenon. It remains an open question, the extent to which non-causal explanation is a heterogeneous kind, including not only the structural model explanations discussed above, but also distinctively mathematical explanations (e.g., Lange 2013), and potentially others as well.⁷ In this section, I review several recent proposals for characterizing non-causal explanations, noting their strengths and weaknesses, in order to extract what I take to be a defensible core conception of non-causal explanation.

Robert Batterman and Rice (2014) have defended a kind of non-causal model-based explanation in terms of what they call “minimal model” explanations. The idea of a minimal model can be traced back to the work of physicists such as Leo Kadanoff and Nigel Goldenfeld in their work on complex phenomena such as phase transitions and the renormalization group (see, e.g., Goldenfeld and Kadanoff 1999). The central idea is that the essential physics of a complex phenomenon can often be captured by a simplistic model that ignores most of the underlying causal details. Batterman and Rice argue that these minimal models, which are found in a wide range of fields

⁶ This expression “physical insight” is the one used by the physicists themselves to describe the advantage of semiclassical explanations over purely quantum ones. It can be further unpacked in terms of the notions of providing true modal information and licensing correct inferences, as above.

⁷ Unfortunately the literature on non-causal explanation is still at the stage of trying to find a core set of examples of non-causal explanation that can be agreed upon. The further task of then trying to create a taxonomy of the different kinds of non-causal explanation still remains to be done.

(including biology), can be used to explain patterns of macroscopic behavior across systems that are quite heterogeneous at smaller scales. In the context of the LGA (Lattice Gas Automaton) minimal model, they explain:

[T]he model is explanatory... because of a backstory about why various details that distinguish fluids... from one another are essentially irrelevant. This delimits the universality class and guarantees a kind of robustness... under rather dramatic changes in the lower-scale makeup of the various systems.... The renormalization group strategy, in delimiting the universality class, provides the relevant modal structure that makes the model explanatory.

(Batterman and Rice 2014: 364)

These simplistic minimal models are explanatory insofar as it can be shown that the minimal model and the realistic system to be explained fall into the same universality class and the model displays the relevant modal structure. There is some confusion in the literature over what exactly is meant by ‘relevant modal structure’ here: On one interpretation, it could just mean what I have discussed above as capturing the relevant patterns of counterfactual dependence in the explanandum phenomenon, a view that I have endorsed. On the other hand, Rice (2015) in particular has emphasized that it should be understood as facts about *independence*, which is an approach that has been criticized by Lina Jansson and Saatsi (forthcoming).⁸

Batterman and Rice go on to argue that these model-based explanations are a non-causal form of explanation, “distinct from various causal, mechanical, difference-making, and so on, strategies prominent in the literature” (Batterman and Rice 2014: 349). They reject the “3M” account of Kaplan and Craver (2011) that requires a mapping between the elements of the model and the actual causal mechanisms. They continue:

Many models are explanatory even though they do not accurately describe the actual causal mechanisms that produced the phenomenon.... [And] there are several reasons why the explanation provided by a model might be improved by removing various details concerning causal mechanisms. (Batterman and Rice 2014: 352)

This is precisely what minimal models do: they ignore the causal details that distinguish the particular different members of a universality class. As Reutlinger (2014) has noted, however, one must be careful in that simply failing to “accurately describe causal mechanisms” and “removing details concerning causal mechanisms” does not automatically mean that one has a non-causal explanation.⁹ As Michael Strevens (2008) has rightly stressed, many causal explanations do this as well.

⁸ This point about an ambiguity in Batterman and Rice’s “modal structure” I owe to Juha Saatsi (personal communication).

⁹ Although Reutlinger takes a weak interpretation of Batterman and Rice’s claims here, and criticizes them for taking this as sufficient for being non-causal, I believe they intend a stronger reading of these claims, which is in fact more in line with the view being defended here. Either way, further clarifications are required. Reutlinger’s views are discussed further below.

Yet another approach to non-causal explanation is Marc Lange's (2013) distinctively mathematical explanations in science. These explanations make use of mathematics, but have as their target physical facts (not mathematical theorems). Not all explanations that make use of mathematics, however, count as distinctively mathematical. Many causal explanations, for example, cite mathematical facts as part of their explanans. Instead, distinctively mathematical explanations are ones where "the facts doing the explaining are modally stronger than ordinary causal laws (since they concern the framework of any possible causal relation)" (Lange 2013: 485). Lange gives as an example of a distinctively mathematical explanation the case of why a parent cannot divide 23 whole strawberries evenly among three children, as being due to the mathematical fact that 23 is not divisible by 3. The explanation depends on the mathematical fact that it is impossible to divide 23 by 3 regardless of the causal entities or processes involved.

Lange argues that distinctively mathematical explanations are a non-causal form of explanation, even though they may include causal information about the explanandum. He writes:

I agree... that distinctively mathematical explanations in science are noncausal. But I do not accept Batterman's ([2010]: 3) diagnosis that what makes these explanations non-causal is that they involve a 'systematic throwing away of various causal and physical details'.

(Lange 2013: 506)

It is not whether or not causal facts are mentioned, or mentioned only very abstractly that characterizes non-causal explanation. Rather, for Lange it is whether the facts doing the explaining are 'more necessary' than ordinary causal laws. While Lange is right to call attention to this question of whether or not the explanation works *by virtue* of citing causal facts, it is not clear that a modally stronger notion of necessity is required for an explanation to count as non-causal.

Yet a third approach to non-causal explanation rejects both Batterman's and Lange's approaches. Reutlinger (2014), like Batterman, defends renormalization group (RG) explanations of universal macro-behavior as a case of non-causal explanation. However, he argues that "Batterman misidentifies the reason that RG explanations are non-causal: he is wrong to claim that if an explanation ignores causal (micro) details, then it is not a causal explanation" (Reutlinger 2014: 1169). As Reutlinger notes, more recent advocates of causal explanation allow that all sorts of irrelevant (non-difference making) causal details can be omitted, without undermining its status as a causal explanation. Reutlinger also disagrees with Lange (2013), however, that what he calls "metaphysical necessity [sic]"¹⁰ is the distinctive characteristic of a non-causal explanation. He writes:

[O]ne need not appeal to metaphysical necessity in order to claim that mathematical facts explain in a noncausal way. All one needs to establish is that the mathematics does not explain by referring to causal facts. (Reutlinger 2014: 1167–8)

¹⁰ It is not clear why Reutlinger switches Lange's "modally stronger" notion of necessity to "metaphysical necessity".

In the context of renormalization group explanations he continues:

RG explanations are noncausal explanations because their explanatory power is due to the application of mathematical operations, which do not serve the purpose of representing causal relations. (Reutlinger 2014: 1169)

The key question here, which I think is roughly right, is whether or not the explanatory factors are a representation of the causal facts and relations. More needs to be said, however, about what is to count as representing causal facts.¹¹ When this is fleshed out, I think Reutlinger and Batterman are in closer agreement than they might realize.

Yet a fourth approach to distinguishing non-causal explanation is given by Lauren Ross (2015), who sheds further light on this question of what it means to *not* be a representation of causal facts. As an example of a non-causal model explanation Ross discusses a dynamical model in neuroscience known as the “canonical” (or Ementrout-Kopell) model. This model is used to explain why diverse neural systems (e.g., rat hippocampal neurons, crustacean motor neurons, and human cortical neurons) all exhibit the same “class I” excitability behavior. She writes:

The canonical model and abstraction techniques used in this approach explain why molecularly diverse neural systems all exhibit the same qualitative behavior and why this behavior is captured in the canonical model. (Ross 2015: 41)

In other words, there are principled mathematical abstraction techniques that show how the detailed models of different neural systems exhibiting class I excitability behavior can all be transformed into the same canonical model exhibiting the behavior of interest. The resulting canonical model is a minimal model in Batterman’s sense.

Ross further argues that these canonical model explanations are a non-causal form of explanation. She writes:

The canonical model approach contrasts with Kaplan and Craver’s claims because it is used to explain the shared behavior of neural systems without revealing their underlying causal mechanical structure. As the neural systems that share this behavior consist of differing causal mechanisms... a mechanistic model that represented the causal structure of any single neural system would no longer represent the entire class of systems with this behavior. (Ross 2015: 46)

It is important to note that not just any abstraction from causal detail makes an explanation non-causal. Rather, it is because the canonical model is able to explain the behavior of neural systems with very different underlying causal-mechanical details—that is, it is an abstraction across very different causal mechanisms—that this model explanation can be counted as non-causal.¹²

¹¹ Reutlinger’s own approach here in (2014) and in (2016) is to deploy what he calls the “folk theory of causation” and the “Russellian criteria” of asymmetry, distinctness of relata, and metaphysical contingency (2014: 1158). While this is an important approach, there are other possible ways one could go about fleshing out what is, or is not, to count as representing causal facts (as will be discussed further below).

¹² I will come back to further elaborate this key idea after introducing the central case of sand ripples.

From these four accounts of non-causal explanation, we can begin to see a convergence towards a core conception of non-causal explanation: A non-causal explanation is one where the explanatory model is decoupled from the different possible kinds of causal mechanisms that could realize the explanandum phenomenon, such that the explanans is not a representation (even an idealized one) of any causal process or mechanism. Before elaborating this core conception of non-causal explanation further, it will be helpful to have a concrete example of a phenomenon for which there is both a causal and a non-causal explanation, to more clearly see how they differ. Such an example is found in the explanandum of how regularly-spaced sand ripples are formed.

4. Explaining the Formation of Sand Ripples

The study of sand ripples belongs to a field known as aeolian geomorphology. Named after the Greek god of wind, Aeolus, aeolian geomorphology is the study of landscapes that are shaped predominantly by the wind, such as the “sand seas” of the Saharan desert, coastal dunes of Namib in southwestern Africa, the Great Sandy Desert of central Australia, the Takla Makan of western China, and the Algodones dunes of southeastern California (see Figure 7.1).



Figure 7.1 A “sand sea”: the Algodones dunes of SE California.
Note the ripples in the foreground, which are superimposed on the dunes.
(Photo courtesy of Eishi Noguchi)

Not only are sand seas (also known as ergs or dune fields) found all over the world, they are also found on other worlds, such as Venus, Mars, and Saturn's moon Titan (the last of which contains the largest sand sea in our solar system at roughly 12–18 km²).

Although wind-blown sand might seem like a simple system, it can organize into vast, strikingly patterned fields, such as the barchan dunes of the Arabian Peninsula's Rub' al Khali that can maintain their characteristic crescent shape and size even while traveling across the desert floor and linking to form a vast filigree pattern. There are different aeolian sand bedforms¹³ that form at different characteristic spatial and temporal scales (e.g., Wilson 1972). At the smallest scale are ripples, which are a series of regular linear crests and troughs, typically spaced a few centimeters apart and formed in minutes. At an even larger scale are dunes, which come in one of a few characteristic shapes (e.g., linear, barchan, star, crescent, or dome); they are typically tens of meters to a kilometer in size and form over years. At the largest scale are draas (also known as megadunes) which are typically 1 km to 6 km in size, and which form over centuries (or even millennia). Interestingly, it is not the case that ripples grow into dunes, or dunes into draas; rather, all three bedforms can be found superimposed at a single site.

The explanandum phenomenon of interest here is the formation of the smallest scale aeolian bedform: sand ripples. Why do sand ripples form an ordered pattern with a particular characteristic wavelength (i.e., a roughly uniform spacing between adjacent crests)? Although it might seem like a straightforward question regarding a simple system, it turns out that answering it is highly nontrivial. There are currently two (different) received explanations in the scientific literature for the formation of regularly spaced sand ripples. The first is a model explanation introduced by Robert Anderson in 1987 (which I will call the "reptation" model explanation of ripples), and the second is a model explanation introduced in 1999 by Brad Werner and Gary Kocurek (which is called the "defect dynamics" model explanation). These two explanations, each of which will be discussed in turn, are not viewed as rivals or competitors, but rather are complementary explanations (a point I will come back to elaborate below). I will argue that while one of them is properly classified as a causal explanation, the other is a non-causal explanation of the formation of ripples.

Anderson's (1987) model explanation marked an important shift in scientists' thinking about the formation of ripples. Since the 1940s it had been assumed that ripples are formed by a barrage of saltating grains of sand, and that the ripple wavelength is determined by the characteristic path length in saltation. Saltation is the process by which a grain of sand gets lifted off the surface, momentarily entrained in the wind, before gravity sends it back down to the surface, typically "splashing" the other grains of sand in the bed before bouncing up again on its next saltation hop. The sand grains that are splashed "creep" forward on shorter, much less energetic trajectories in a process called reptation. The processes of saltation and reptation are depicted in Figure 7.2.

¹³ A 'bedform' is a generic term in the geosciences for "pile of stuff", and in the context of aeolian geomorphology it typically means a pile of sand, such as a ripple or sand dune.

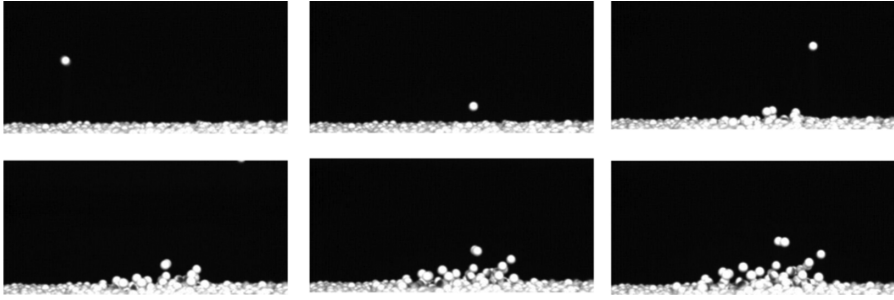


Figure 7.2 A sequence of high-speed motion photographs of the processes of saltation and reptation.

Note the energetic saltation particle coming in from upper left in the first frame is already on its way (after its bounce) to its next hop by the third frame. The particles in the bed that were splashed by the impact of the saltating particle creep forward (but do not rebound) in the process of reptation.

(From Beladjine et al. 2007: Fig. 2)

In his pioneering 1941 book, *The Physics of Blown Sand and Sand Dunes*, Ralph Bagnold hypothesized that the key causal process in the formation of ripples of a particular wavelength is saltation. Bagnold writes:

This remarkable agreement between the range, as calculated theoretically... and the wavelength of the real ripples, suggest strongly that the latter is indeed a physical manifestation of the length of the hop made by the average sand grain in its journey down-wind.

(Bagnold [1941] 2005: 64)

This hypothesis ran into several difficulties, however. One of the distinctive features of ripple formation is that the ripples begin close together and then grow in wavelength before reaching a stable characteristic spacing. Even by the 1960s it was realized that “[t]here can be no question about the progressive growth and increase in size of the ripples... [and it] is difficult to reconcile with Bagnold’s concept of a characteristic path length” (Sharp 1963: 628). It was not until the late 1980s that an acceptable model explanation that could accommodate this feature was formulated.

Anderson agrees with Bagnold that ripple formation is *not* the direct result of fluid forces imposed by the air (Anderson 1987: 944). Unlike Bagnold, however, Anderson identifies reptation as the key causal process in the formation of ripples and argues that saltating grains makes a negligible contribution to ripples. The way in which reptation comes in to explain ripple formation, however, is not as straightforward as one might have hoped. Rather than trying to track the trajectories and forces acting on every grain of sand, Anderson explains the growth and spacing of ripples using an idealized model. This numerical model shows how a seemingly random barrage of reptating grains of sand can surprisingly lead to the emergence of a dominant characteristic wavelength for the ripples.

Anderson’s model explanation makes a number of idealizing assumptions. First, the grain-bed interaction is characterized statistically in terms of a “splash function” that,

for a given distribution of impact velocities, gives the number of ejected grains and a probability distribution for their ejection velocities. Second, the wide distribution of actual trajectories is idealized to two end members: high energy successive saltations and low-energy reptations, such that “the successive saltation population has zero probability of death [the bounces always perfectly reproduce themselves, never decaying] and the reptations have exactly *unit* probability of death upon impact [they neither reproduce themselves nor give ‘birth’ to other trajectories]” (Anderson 1987: 947). Third, it assumed that the spatial distribution of saltation impacts on a horizontal surface is uniform, and that they all descend at an identical angle. Fourth, the low number of grains traveling in high energy trajectories, and the low probability they will be incorporated into the ripple bed,

allows us to ignore their *direct* contribution to ripple transport. Rather, their role in ripple formation and translation is here idealized as merely an energy supply for initiating and maintaining reptation (Anderson 1987: 947)

Here we see the shift to the view that reptation—not saltation—is the key process in ripple formation, and saltation is simply a generic energy source for reptation. Additionally, the role of wind shear stresses is neglected and it is assumed that the bed is composed of identical grains of sand (this latter assumption is reasonable for what are known as ‘well-sorted’ aeolian sands in places like the Sahara, but fails for places with bimodal or poorly sorted sand).

With these idealizing assumptions, Anderson introduces the following numerical model of the sand flux as a function of position (Anderson 1987: 951).

$$Q(x) = Q_0 + q_{ej} \cot \alpha \int_0^{\infty} [z(x) - z(x-a)] p(a) da. \quad (1)$$

The first term in Equation (1), Q_0 , represents the total expected mass flux across the bed due to both saltation and reptation; the second term represents the spatially varying flux due to the growth and movement of ripples. More specifically, q_{ej} is the mass ejection rate, α is the incident angle of the impacting grains, z is the bed elevation, and $p(a)da$ is the probability distribution of the different reptation lengths. One can then use this equation, along with the sediment continuity equation and expression for bed elevation, to obtain the growth rate and translation speeds of bed perturbations of various wavelengths.

If one considers a reasonably realistic exponential or gamma probability function for the reptation lengths, and then performs a Fourier transform, these yield the dimensionless real and imaginary components of the phase speed. Anderson summarizes the results of this analysis as follows:

The most striking alteration of the pattern of ripple growth resulting from the introduction of [these] more realistic probability distributions of reptation lengths is the dampening of the growth of the shorter wavelength harmonics.... [T]here exists a single fastest-growing

wavenumber corresponding to wavelengths on the order of six times the mean reptation length for both the exponential and gamma distributions. (Anderson 1987: 953)

In other words, this model shows how a seemingly random splashing of sand grains can lead to the formation of ripples with a specific characteristic wavelength. Although this analysis vindicates the view that ripple wavelength is controlled by the process of reptation not saltation, Anderson is careful to note that the relation is not one of

a simple equivalence between transport distance and ripple length. The relevant physics is not a rhythmic barrage of trajectories of length equal to the ripple spacing; it is a pattern of divergence and convergence of mass flux dominated by reptating grains with a probability distribution of reptation lengths. (Anderson 1987: 955)

Not only do observations in nature and wind-tunnel experiments agree reasonably well with the wavelength predicted by this model, but the model also captures the way in which ripple spacing varies with changes in the mean reptation distance under different conditions.

How should we classify this model-based explanation? It is worth pausing to summarize some of the key features of this explanation. First, as we saw, this explanation omits many causal details (e.g., wind shear stresses, the contribution of saltation particles to ripples, etc.). Second, we have a statistical characterization of key processes (e.g., the ‘splash function’ for grain-bed interaction). Third, the explanation involves many idealizations (e.g., about the allowed kinetic energies and angles of the trajectories). Fourth, it involves highly mathematical models and analyses (e.g., complex phase speeds, Fourier transforms, etc.). Nonetheless, I argue that it is still a causal explanation. This is because the mathematics and model explanation are still a straightforward and direct representation of the relevant fundamental causal processes, causal mechanisms, and causal entities that we know to be operating in that domain. An incomplete, idealized, and statistical representation of a causal process is still a representation of a causal process.

It is helpful to recall that a mathematical model in science really consists of two models. First, there is what is called the *conceptual model*, which is a conceptualization or ‘picture’ of what is going on in the system.¹⁴ It is a particular conception about what the relevant entities, processes, and interactions are in a particular domain, *prior to* any particular mathematical (or physical) representation of those entities, processes, and interactions. Second, there is the choice of a particular mathematical representation of that conceptual model. There are different possible mathematical representations for one and the same conceptual model, and one and the same mathematical model can be used to represent different conceptual models—even of different physical systems, such as in the case of physical analogies of the sort exploited by James Clerk Maxwell (see Bokulich 2015 for a discussion). So when a model fails, one can

¹⁴ For a historical discussion of this distinction between conceptual models and mathematical models see Bokulich and Oreskes (2017: Section 41.2).

ask whether it was due to an inadequate conceptual model or due to an inadequate mathematical representation of that conceptual model, or both. In the case of Anderson's (1987) model of ripple formation, the underlying conceptual model is a causal one. It is a causal model despite its idealized and mathematical character because the mathematics is still a direct representation of the basic causal entities and causal processes in that physical system.

The second model-based explanation of the formation of ripples, due to Werner and Kocurek (1999), is a different story. Instead of formulating the explanation in terms of the relevant fundamental entities and causal processes (e.g., the saltation and reptation of grains of sand moving under the force of gravity, etc.), this second explanation introduces a new "pseudo-ontology" at the more abstract level of bedform structures, and makes them the dynamical variables through which the system evolves and the phenomenon is explained. The pseudo-ontology they introduce is that of a pattern "defect", and the model describes how these defects dynamically evolve and interact over time to produce the regular spacing of ripples.

A defect is defined most broadly as an imperfection in a pattern.¹⁵ Conceptually, one works backwards from the end-state of a perfectly ordered set of parallel ripples, of uniform height and uniform spacing, whose crest lines span the entire width of the bedform. One kind of defect, called a "termination", is an interruption or break in the crest line. When there are two opposite-facing free ends, these are referred to as a termination and anti-termination pair. A crest line with only one break would have

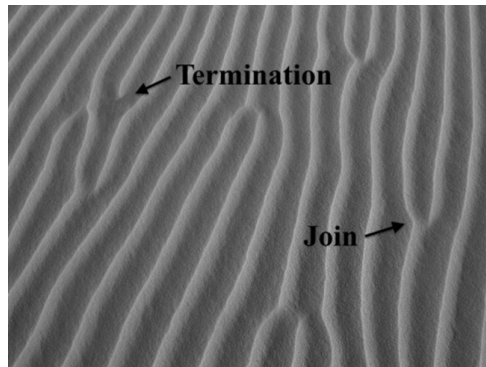


Figure 7.3 Examples of ripple defects.

As suggested by this photograph, terminations can propagate downwind, joining the next ripple crest ahead of it, becoming a join temporarily before breaking off again on the other side of the ripple.

¹⁵ When the discussion of defects was first introduced into geomorphology, an analogy was explicitly made to defects in material science, such as in the case of dislocations or defects in a crystal lattice (Anderson and McDonald 1990: 1344). While one might think that defects are unimportant, the presence of defects in a crystal lattice, for example, can have a tremendous effect on the physical properties of the crystal (see Lifshitz and Kosevich 1966 for a review).

a low density of defects, while a crest line with many breaks would have a high density of defects. Another kind of defect is known as a “join” (or “bifurcation”), where two crest lines, instead of being parallel, form a Y-junction. These two key types of defects are depicted in Figure 7.3.

An aeolian bedform starts out in a largely disordered state with a high density of defects. The crest lines are short, being interrupted by many terminations, and adjacent crest lines begin close together. Detailed field observations show that as these defects become eliminated (e.g., by termination/anti-termination pairs meeting up to form a longer continuous ripple crest line), the spacing between adjacent crest lines (the wavelength) grows rapidly at first, and then slows down over time until the final characteristic wavelength of ordered bedform of ripples is reached.

Rather than analyzing this process of ripple formation at the scale of grains of sand that are reptating, the approach of the defect dynamics model explanation is to couple spacing and number of defects as the relevant dynamical variables. Kocurek and colleagues argue that the other “explanation for these patterns . . . is that they are self-organized. . . the proposal is that it is the interactions between the bedforms themselves that give rise to the field-scale pattern” (Kocurek et al. 2010: 51). They elaborate on this alternative as follows:

The self-organization hypothesis represents an alternative explanation to reductionism, in which large-scale processes such as bedform-pattern development are thought to arise as the summation of smaller-scale processes (e.g., the nature of grain transport causes the spacing pattern in wind ripples). (Kocurek et al. 2010: 52)

Although philosophers of science typically use the term ‘reductionism’ in a slightly different way, it is clear in these quotations that the defect dynamics explanation is, first, seen as an *alternative* to the reptation model explanation of ripples, and second, seen as an explanation that is *not* a causal story about how grain transport causes the formation of ripples. To understand and assess these two claims, we must take a closer look at the defect dynamics model explanation.

As with Anderson (1987), the explanation is an idealized model-based explanation. The defect dynamics explanation exploits the geometrical properties of an idealized representation of a bedform field with ripple crests and defects. Suppose the bedform field where the ripples form is of width X and length Y . In the limit of a perfectly ordered bedform field, where all the ripple crest lines are continuous across the entire width, X , of the field, and have achieved their final characteristic spacing (or wavelength) λ , then the total (possible) crest length is given by

$$L = XY / \lambda \quad (2)$$

where the total number of ripples (crest lines of length X) is Y / λ . The two variables being tracked over time are the *mean spacing* between bedforms,

$$\lambda(t) = A / L, \quad (3)$$

and the *defect density*,

$$\rho(t) = N / L, \quad (4)$$

which is the number of defect pairs (terminations and anti-terminations) per unit length of crest line.

As ripples are forming, they translate downwind, in a direction normal to the orientation of the crest line. In order to describe the evolution of the system at this level, one needs to define the mean velocity, v_b at which the bedforms (ripples) migrate:

$$v_b = \gamma / \lambda \quad (5)$$

where γ is equal to the sediment flux times the bedform index (the ratio of spacing to height, which is assumed to be constant).¹⁶ The other relevant “entity” in this model explanation is the defect, which migrates at a mean velocity, v_d , that is roughly three times the mean velocity of the bedform, $v_d = \alpha v_b$ (Werner and Kocurek 1997: 772). The defects migrate faster than the ripples, because the crest line of a termination is shorter than a full ripple and the termination involves a tapering of the ripple height down to zero; intuitively, they move faster simply because there is less sand to move. If you were to watch this process unfold, you would see the defects (the broken end “termination” of a crest line) propagate towards the crest line ahead of it, meet up with that crest line to form a join (Y-junction), before the downwind branch of the Y-junction breaks off, then starts to propagate towards the crest line ahead of it; it then forms another Y-junction again, and the process repeats. The overall appearance is of a single defect passing through successive ripples as it propagates more rapidly downwind.¹⁷

Each time a defect passes through a bedform crest, it loses a small segment, l_0 , of its length, because smaller bedforms tend to merge or get absorbed by larger bedforms. This results in a (slower) lateral movement as well: leftward for terminations and rightward for anti-terminations. So far we have defects propagating rapidly downwind and slowly towards the outside edges of the ripple field. The process by which the defects get eliminated, and the field progresses from a disordered state to a highly ordered state of continuous, uniformly spaced ripples is as follows: when a left facing termination (in its downwind and lateral movement) encounters an anti-termination, the two defects “annihilate” forming a stable continuous crest line. If a defect does not encounter its anti-termination “pair”, then it eventually gets eliminated at the boundary of the field when it runs out of sand. Using the general geometrical constraints and formulating these processes in terms of the time rate of change for the total crest length, L , and the time rate of change of the number of defect pairs (understood as the sum of the rates of both pair annihilation and boundary annihilation) and expressing

¹⁶ The presentation here follows Werner and Kocurek (1999) and (1997), where further details can be found.

¹⁷ Although the defect looks like a single unified thing, maintaining its identity as it moves continuously through space and time, the sand that makes up that defect is continuously changing.

these in terms of the variables of defect density, ρ , and mean spacing, λ , leads to the following set of coupled, nonlinear differential equations:¹⁸

$$\frac{d\lambda}{dt} = -2 \frac{dL_d}{dt} = \rho\lambda \quad (6)$$

$$\frac{d\rho}{dt} = -r|v_d - v_b|\rho^2 + \left(\frac{dL_d}{dt} - r|v_d - v_b| \right) \frac{\rho}{X} - |v_d - v_b| \frac{\rho}{Y} \quad (7)$$

$$\frac{dL_d}{dt} = -\frac{\gamma|\alpha - 1|l_0}{\lambda^2}, \quad |v_d - v_b| = \frac{\gamma|\alpha - 1|}{\lambda} \quad (8)$$

We can see why the spacing, λ , grows rapidly at first when there are lots of defects, but then as the defect density goes down, there are fewer opportunities for crest length to become reduced. This means that the total crest length, L , will asymptotically approach some value, which because of the fixed area, $A = XY$, means in turn that the wavelength (mean spacing) $\lambda = XY / L$ will also change more slowly as it approaches a fixed value.

The defect dynamics explanation, like Anderson's reptation model explanation, is able to produce realistic spacing values for ripples that match observations, and moreover, is able to explain in a very intuitive way how and why that spacing changes over time in the way that it does. How should this model explanation be classified? Werner and Kocurek (1999: 727) argue that what distinguishes the defect dynamics explanation is that it "permits a treatment that bypasses fundamental mechanisms". In other words, they do not see this explanation as working by citing the causal processes involved. Indeed they argue that the fact that this explanation can work despite ignoring the operative causal processes "call[s] into question the widespread assumption that bed-form spacing approaches a steady-state value characteristic of fluid flow and sediment transport" (Werner and Kocurek 1999: 727), where fluid flow (wind) and sediment transport (saltation and reptation) are clearly the relevant fundamental causal processes in this system. One might worry that *pace* Werner and Kocurek, the defect dynamics explanation really is an explanation in terms of those fundamental causal mechanisms, just those causal mechanisms described at a higher, perhaps aggregated level. As long as it was still those particular causal process (e.g., reptation) that were grounding the force of the explanation, or as I prefer to put it, if the defect explanation was still a straightforward representation of those causal processes, then it would still count as a causal explanation. To see why this is not the case, however, one more feature of the defect dynamics explanation must be explored.

It turns out that the defect dynamics explanation is not just an explanation for the formation of aeolian (wind) ripples, but it is also an explanation for the formation of subaqueous (underwater) ripples (Figure 7.4).

¹⁸ Further details in deriving these equations can be found in Werner and Kocurek (1999).

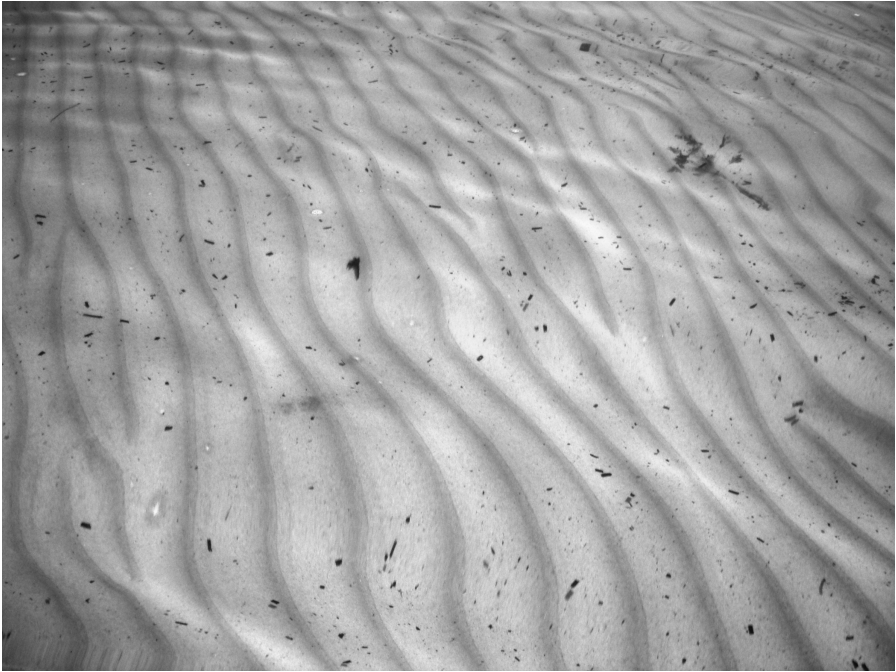


Figure 7.4 Subaqueous sand ripples on the ocean floor.

Note the presence of pattern defects, such as the join (in the upper left), and the termination and anti-termination (in the center right).

Although the patterns that these two systems form are the same, the causal mechanisms by which they form are completely different. Recall that in the case of aeolian ripples it was the bombardment by saltating grains of sand that “splashed” into the bed, causing the other grains to reptate. In the case of subaqueous ripples, however, because of the greater density of water, saltating grains of sand impact the bed too feebly to cause either continued saltation or the reptation of other grains. Reptation is not a relevant causal process in the formation of subaqueous ripples. Similarly, while wind-shear stresses were completely negligible in the case of aeolian ripples, in the case of subaqueous ripples, bottom shear stress due to fluid flow is all important, being what directly transports each grain of sand. This important difference was recognized early on by Bagnold who writes:

That too great a reliance on a similarity of effect as an indication of a similarity of cause may lead to a confusion of ideas, is well exemplified by the case of sand ripples. Everyone is familiar with the pattern of sand ripples on a sea beach. . . . And it would be hard indeed to find a single point wherein they differ in appearance from the wind ripples seen on the surfaces of dunes. Yet the mechanism of their formation cannot be the same in the two cases. The conditions are quite different. The beach ripple is due essentially to the *alternating* flow of water backwards and forwards under successive wavelets. (Bagnold [1941] 2005: 162 emphasis original)

Despite the very different causal explanations for aeolian and subaqueous sand ripples, they both can be equally well explained by the defect dynamics model explanation. In the subaqueous case, the formation of a well-ordered ripple field of a particular wavelength is also explained by the more rapid propagation of defects through the crests and their annihilation upon encountering an anti-termination pair.

The defect dynamics explanation is, I argue, a *non-causal* explanation. This is not because it is an idealized representation that leaves out many details, nor is it because it involves a characterization of the phenomenon in terms of a highly mathematical model. Rather, it is because the mathematics is not a representation of a conceptual model about the relevant causal processes operating in that system. If we were to take a step back and ask any geoscientist today: What are the relevant causal entities and causal processes involved in the formation of aeolian ripples? The answer would be grains of sand undergoing saltation (initiated by wind, and propelled by gravity) and grains of sand undergoing reptation (due to the splash-down impact, where a little of that kinetic energy is distributed among a much larger number of grains of sand). While Anderson's model explanation is a mathematical representation of a conceptual model about these causal processes, the defect dynamics model is not. Similarly, if one were to ask what are the causal processes involved in the subaqueous ripples case, the answer would clearly not be saltation and reptation, which do not occur in this system, but rather fluid shear stresses in an alternating current, directly transporting grains of sand (a different set of causal processes).

While Anderson's (1987) model explanation is an explanation of the formation of aeolian ripples, it is *not* an explanation of the formation of subaqueous ripples. In representing the causal processes involved in the aeolian case, it *cannot* also represent the (different) causal processes in the subaqueous case. They are fundamentally different *types* of causal processes (not merely different token causal processes of the same type causal process, the latter of which could be accommodated by the same causal model explanation). The fact that the defect dynamics model explanation is an explanation of *both* the formation of aeolian ripples *and* the formation of subaqueous ripples makes clear that it is not a representation of the causal processes at all.

5. Conclusion

The question of what it means to be a non-causal explanation turns out to be a subtle issue. Although the different proposals reviewed in section 3 were *prima facie* disagreeing with one another, I argued that they could each be interpreted as orbiting what I take to be a common core conception of non-causal explanation.¹⁹ Moreover,

¹⁹ While there may be forms of non-causal explanation that fall outside of this core conception (such as perhaps Lange's distinctively mathematical explanation), this core conception nonetheless is able to capture some of the key features common to many of the examples of non-causal explanation discussed in the literature.

I argued that this core conception is also exemplified by the defect dynamics explanation of the formation of ripples, discussed above. As with Batterman and Rice's (2014) examples, ripple pattern formation can be understood as a kind of universal phenomenon that is realized by diverse causal systems.²⁰ While there is a sense in which the formation of the ripple pattern is "modally stronger", as Lange (2013) puts it, than the particular causal laws that realize it in the aeolian case, for example, it is not clear that Reutlinger's (2014) "metaphysical necessity" is the right way to describe this. As Reutlinger (2014) rightly notes, however, a non-causal explanation is one where the mathematical model does not serve the purpose of representing the causal processes, and as Ross (2015) further emphasizes, it is a model explanation that is abstracted across *different types* of causal processes and mechanisms. To reiterate, a non-causal explanation is one where the explanatory model is decoupled from the different possible kinds of causal mechanisms that could realize the explanandum phenomenon, such that the explanans is not a representation (even an idealized one) of any causal process or mechanism.²¹

To say that a particular explanation is non-causal does not entail that the explanandum is a purely mathematical phenomenon. The defect dynamics model explanation is a non-causal explanation of a *physical* phenomenon: the formation of real sand ripples. The defect dynamics explanation simply has the further advantage that it can be applied not only to aeolian ripples, but also to subaqueous ripples. Moreover, to say that these physical phenomena have a non-causal explanation does not mean that they are somehow "uncaused" events. In both the aeolian and subaqueous ripple cases, there is no doubt that there is a complete causal story (or more precisely two different complete causal stories) to be told about the formation of these ripples. As we saw in detail for the aeolian case, we even have such a causal explanation in hand.

The existence of a causal explanation does nothing to undermine the explanatory value of a non-causal explanation. As Holly Andersen (forthcoming) has cogently argued, there are many different ways in which causal and non-causal (or what she calls mathematical) explanations can be complementary. The reptation model explanation and the defect dynamics model explanation are not rivals. Each type of explanation serves to bring out different features of the phenomenon more clearly and offers different sorts of insights into its nature. This is what I earlier described as type II explanatory pluralism: there can be more than one scientifically acceptable explanation for a given phenomenon at a time. One could even go further and argue that while

²⁰ It is in fact even more universal than I have discussed here, being applicable not only to aeolian and subaqueous sand ripples, but also systems of sand bars, what are called 'sorted bedforms' (an underwater sorting of grains of different sizes), and linear dunes, which occur both here on Earth and elsewhere, such as on Titan where there are very different grain, atmospheric, and gravitational conditions.

²¹ Although universal phenomena are a natural place to look for non-causal explanations, not all non-causal explanations need involve universality. The non-causal semiclassical explanations of quantum phenomena, such as wavefunction scarring, are a case in point: although they do not involve universality, they do satisfy this definition insofar as they are not a direct representation of the causal entities or processes operating in that system (indeed the entities deployed in the semiclassical explanation are fictions).

there are some respects in which the reptation model explanation is deeper than the defect dynamics model explanation, there are other respects in which the defects explanation can be seen as deeper than the reptation explanation.²² This pluralism, rather than revealing some sort of shortcoming in our understanding of sand ripples, is in fact one of its great strengths.

The analysis presented here suggests that non-causal explanations may not in fact be as rare or strange as they have hitherto been assumed to be. We are increasingly learning that universal phenomena, across fundamentally different types of causal systems, are widespread among the sciences (whether it is phase transitions in different substances, class I excitability in diverse neural systems, or ripple formation in different environments). The defect dynamics model explanation of ripple formation is able to account for this universality by decoupling the explanation from the particular types of causal stories that might realize it. It is not because the model explanation is idealized, leaves out many causal details, or because it is formulated in terms of an abstract mathematical model, that makes it non-causal. The defect dynamics explanation is non-causal because it is not a representation of the causal processes at all. If it were a representation of the causal processes occurring, for example, in the case of aeolian ripples, then it could not also be an explanation for the formation of subaqueous ripples, and vice versa. Moreover, the fact that we can give a causal explanation in the aeolian ripple case does not rule out there being a scientifically accepted non-causal explanation of aeolian ripples as well. As the defect dynamics model explanation teaches us, we can indeed find non-causal explanations in a (sand-) sea of causes.

Acknowledgments

I would like to express my deep gratitude to Gary Kocurek for very helpful discussions about aeolian geomorphology and defect dynamics. I am also grateful to the editors for providing helpful feedback on this chapter. Any mistakes are of course my own.

References

- Andersen, H. (forthcoming), 'Complements, Not Competitors: Causal and Mathematical Explanations', *British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axw023.
- Anderson, R. (1987), 'A Theoretical Model for Aeolian Impact Ripples', *Sedimentology* 34: 943–56.
- Anderson, R. and McDonald, R. (1990), 'Bifurcations and Terminations in Eolian Ripples', *Eos* 71: 1344.
- Ayrton, H. ([1904] 1910), 'The Origin and Growth of Ripple-Mark', *Proceedings of the Royal Society of London. Series A: Containing Papers of a Mathematical and Physical Character* 84: 285–310.

²² For a discussion of the different possible dimensions along which explanatory depth can be measured see Hitchcock and Woodward (2003).

- Bagnold, R. ([1941] 2005), *The Physics of Blown Sand and Desert Dunes* (New York: Dover).
- Batterman, R. (2010), 'On the Explanatory Role of Mathematics in Empirical Science', *British Journal for the Philosophy of Science* 61: 1–25.
- Batterman, R. and Rice, C. (2014), 'Minimal Model Explanations', *Philosophy of Science* 81: 349–76.
- Beladjine, D., Ammi, M., Oger, L., and Valance, A. (2007), 'Collision Process between an Incident Bead and a Three-Dimensional Granular Packing', *Physical Review E* 75: 061305, 1–12.
- Bokulich, A. (2008a), *Reexamining the Quantum-Classical Relation: Beyond Reductionism and Pluralism* (Cambridge: Cambridge University Press).
- Bokulich, A. (2008b), 'Can Classical Structures Explain Quantum Phenomena?', *British Journal for the Philosophy of Science* 59: 217–35.
- Bokulich, A. (2011), 'How Scientific Models Can Explain', *Synthese* 180: 33–45.
- Bokulich, A. (2015), 'Maxwell, Helmholtz, and the Unreasonable Effectiveness of the Method of Physical Analogy', *Studies in History and Philosophy of Science* 50: 28–37.
- Bokulich, A. (2016), 'Fiction as a Vehicle for Truth: Moving Beyond the Ontic Conception', *The Monist* 99: 260–79.
- Bokulich, A. (forthcoming), 'Representing and Explaining: The Eikonic Conception of Scientific Explanation.' *Philosophy of Science* (Proceedings).
- Bokulich, A. and Oreskes, N. (2017), 'Models in Geosciences', in L. Magnani and T. Berlotti (eds.), *Springer Handbook of Model-Based Science* (Dordrecht: Springer), 891–912.
- Craver, C. (2007), *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience* (Oxford: Oxford University Press).
- Craver, C. (2014), 'The Ontic Account of Scientific Explanation', in M. I. Kaiser, O. R. Scholz, D. Plenge, and A. Hüttemann (eds.), *Explanation in the Special Sciences: The Case of Biology and History* (Dordrecht: Springer), 27–52.
- Goldenfeld, N. and Kadanoff, L. (1999), 'Simple Lessons from Complexity', *Science* 284: 87–9.
- Hitchcock, C. and Woodward, J. (2003), 'Explanatory Generalizations, Part II: Plumbing Explanatory Depth', *Noûs* 37: 181–99.
- Jansson, L. and Saatsi, J. (forthcoming), 'Explanatory Abstractions', *British Journal for the Philosophy of Science*.
- Kaplan, D. and Craver, C. (2011), 'The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective', *Philosophy of Science* 78: 601–27.
- Kocurek, G., Ewing, R., and Mohrig, D. (2010), 'How do Bedform Patterns Arise? New Views on the Role of Bedform Interactions within a Set of Boundary Conditions', *Earth Surface Processes and Landforms* 35: 51–63.
- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Lewis, D. (1986), 'Causal Explanation', in *Philosophical Papers*, vol. II (New York: Oxford University Press), 214–40.
- Lifshitz, I. and Kosevich, A. (1966), 'The Dynamics of a Crystal Lattice with Defects', *Reports on Progress in Physics* 29 (Part I): 217–54.
- Reutlinger, A. (2014), 'Why Is There Universal Macro-Behavior? Renormalization Group Explanation as Non-Causal Explanation', *Philosophy of Science* 81: 1157–70.

- Reutlinger, A. (2016), 'Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation', *Philosophy of Science* 83: 733–45.
- Rice, C. (2015), 'Moving Beyond Causes: Optimality Models and Scientific Explanation', *Noûs* 49: 589–615.
- Ross, L. (2015), 'Dynamical Models and Explanation in Neuroscience', *Philosophy of Science* 82: 32–54.
- Saatsi, J. (forthcoming), 'On Explanations from "Geometry of Motion"', *British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axw007.
- Saatsi, J. and Pexton, M. (2013), 'Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Non-Causal Explanations', *Philosophy of Science* 80: 613–24.
- Salmon, W. (1984), *Scientific Explanation and the Causal Structure of the World* (Princeton: Princeton University Press).
- Salmon, W. (1989), *Four Decades of Scientific Explanation* (Minneapolis: University of Minnesota Press).
- Sharp, R. (1963), 'Wind Ripples', *Journal of Geology* 71: 617–36.
- Skow, B. (2014), 'Are There Non-Causal Explanations (of Particular Events)?', *British Journal for the Philosophy of Science* 65: 445–67.
- Strevens, M. (2008), *Depth: An Account of Scientific Explanation* (Cambridge, MA: Harvard University Press).
- Werner, B. and Kocurek, G. (1997), 'Bedform Dynamics: Does the Tail Wag the Dog?', *Geology* 25: 771–4.
- Werner, B. and Kocurek, G. (1999), 'Bedform Spacing from Defect Dynamics', *Geology* 27: 727–30.
- Wilson, I. (1972), 'Aeolian Bedforms: Their Development and Origins', *Sedimentology* 19: 173–210.
- Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation* (Oxford: Oxford University Press).

8

The Development and Application of Efficient Coding Explanation in Neuroscience

Mazviita Chirimuuta

1. Introduction

Recent philosophy of neuroscience has been dominated by discussion of mechanisms. The central proposal of work in this tradition is that explanations of the brain are crafted through the discovery and representation of mechanisms. Another core commitment is to explanation being a matter of situating phenomena in the causal structure of the world. This is often accompanied by commitment to an interventionist theory of causation and causal explanation. Accordingly, a criterion of explanatory sufficiency is the ability of a theory or model to tell us how our phenomenon would be altered under different counterfactual scenarios—the ability to answer *what-if-things-had-been-different* or w-questions (Woodward 2003).

Various authors believe that it is useful to decouple the counterfactualist parts of Woodward’s account of explanation from the causal, interventionist ones and thereby develop an account of non-causal explanation.¹ One thing that might seem puzzling about this move is that it extends Woodward’s framework in such a way as to apparently divorce scientific explanation from the demands of working out how to intervene successfully in the world. The tight connection between causally explaining and making a difference was originally one of the selling points of Woodward’s account. Yet if an explanation fulfills the counterfactualist, but not the interventionist norms, it can seem hard to find a point to the investigation beyond theoretical speculation. For when one learns of a non-causal explanation of, say, patterns of spiking and non-spiking activity in a neuron, one is not thereby learning of the specific “levers and pulleys” which

¹ E.g., Bokulich (2011) and Saatsi and Pexton (2013).

would allow one to impede a pathological kind of neuronal behavior, such as underlies epileptic disease.²

I have recently argued that the w-question criterion can be satisfied by models of neural systems which are non-mechanistic (Chirimuuta 2014) and non-causal (Chirimujuta 2017). I refer to these as *efficient coding explanations*. Such explanations occur frequently in computational neuroscience—a broad research area which uses applied mathematics and computer science to model neural systems. The models in question ignore biophysical specifics in order to describe the information processing capacity of a neuron or neuronal population. Such models figure prominently in explanations of *why* a particular neural system exhibits a characteristic behavior. Neuroscientists formulate hypotheses as to the behavior's role in a specific information-processing task, and then show that the observed behavior conforms to (or is consistent with) a theoretically derived prediction about how that information could efficiently be transmitted or encoded in the system, given limited energy resources. They do not involve decomposition of biophysical mechanisms thought to underlie the behavior in question; rather, they take an observed behavior and formulate an explanatory hypothesis about its functional utility. As Doi et al. (2012: 16256) write:

It has been hypothesized that the early stages of sensory processing have evolved to accurately encode environmental signals with the minimal consumption of biological resources.... This theoretical hypothesis, generally known as efficient coding, has been used to explain a variety of observed properties of sensory systems.³

In this chapter I argue that efficient coding explanations have important roles to play in various kinds of practical activity. There are more ways to make a difference than facilitating and preventing causal effects; one may also wish to build things. There is a close and historically embedded connection between engineering and the research traditions in neuroscience which employ efficient coding reasoning.⁴ Thus we find numerous instances of efficient coding reasoning in attempts both to reverse engineer the nervous system and to forward engineer devices which replicate some of the functions of the biological brain. Before discussing these applications, in section 2 I will outline my criteria for non-mechanistic and non-causal explanation, and this will be followed by a case study of explanations of *lateral inhibition* in the early visual system.

² I thank Anna Alexandrova for raising this issue. Even though the interventionist theory of causation only need refer to hypothetical interventions, not actual ones, advocates of interventionism often highlight the connection between this way of thinking about causation and the practice of figuring out ways to alter the course of natural events. E.g., Kaplan and Craver (2011: 602).

³ Efficient coding explanations do not rely on the strong adaptationist assumption that the brain of humans, or any other animal, is optimal. Instead, the point is to show that an observed feature has similarities with a theoretically predicted optimum, though there may be substantial departures from optimality.

⁴ For more on the historical links, see Husbands and Holland (2008) on the Ratio Club (1949–58).

2. Efficient Coding Explanation and the Causal/Non-Causal Frontier

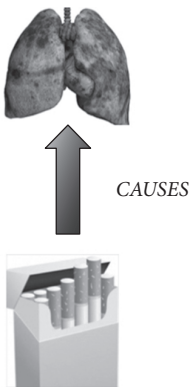
Holly Andersen offers many useful reflections on the much contested frontier between causal and non-causal explanation. Causal explanation is often defined broadly as the placing of the explanandum phenomenon within the network of causal relationships in the world. A more stringent definition asserts that for an explanation to be causal, the connection between the explanans and explanandum must be a causal one (Andersen 2016). This would rule out constitutive mechanistic explanation, since in those cases the relationship between the entities and activities of the explanans, and the explanandum phenomenon, is one of constitution rather than causation. This strikes me as a problematic feature of Andersen's narrow definition. As I see it, constitutive mechanistic explanations where both explanans and explanandum are characterized as a set of causal relationships, should count as a kind of causal explanation. The important point is that the explanandum is *doing* something which brings about the explanans. As Kaplan and Craver (2011: 611) put it, it is important to see how the mechanism “produces, maintains, or underlies the phenomenon”.

The lesson here is that there *is* a difference worth marking between mechanistic and aetiological explanation, but that does not mean that mechanistic explanation is non-causal—it is simply a different kind of causal explanation. By focusing on the relationship between explanans and explanandum we can chart these and other kinds. In each of the four examples depicted in Figure 8.1, the explanandum is a biological phenomenon. In the cases of (a) aetiological and (b) mechanistic explanation, the explanantia are also phenomena which can naturally be described as a series of causal processes. I classify these as two species of causal explanation. In (c) we have the non-causal, mathematical explanation of the hexagonal shape of honeycomb. The explanans is a law or fact of mathematics—the honeycomb conjecture—rather than an empirically observable causal process. One cannot speak of the mathematical facts as causing anything to happen in nature, though they do *constrain* the sequence of biological events.

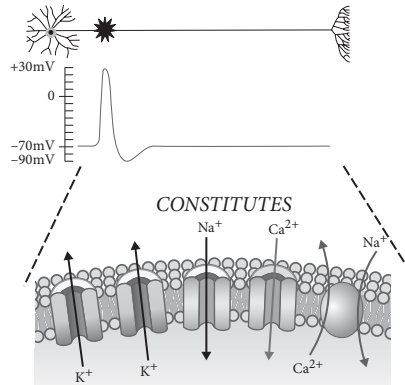
The fourth kind, (d) efficient coding explanation, is clearly different from mechanistic and aetiological explanation in that the explanans is an abstract coding scheme or algorithm, rather than an empirically observable causal process. Also, the relationship between explanans and explanandum is one of *implementation*⁵ rather than causation or constitution. Thus I classify it as a kind of non-aetiological and non-mechanistic explanation. In the cases of efficient coding explanation that I will discuss in this chapter, the neural system is said to implement a specific code or coding strategy, and this reasoning yields insights into why the system behaves in the ways observed.

⁵ For the purposes of this chapter I will bracket the vexed philosophical debate over the proper analysis of this term, noting that the concept of implementation is employed widely within neuroscience. But see Sprevak (2012) for an excellent discussion of the philosophical issues.

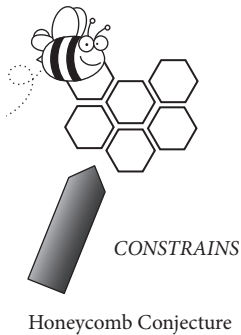
(a) Aetiological Explanation



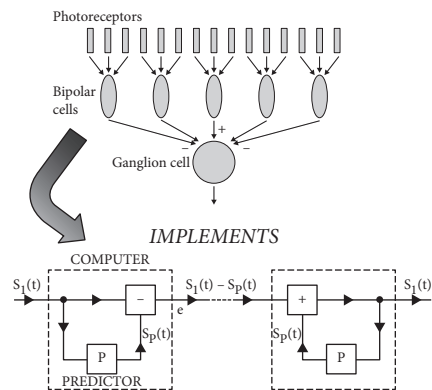
(b) Mechanistic Explanation



(c) Mathematical Explanation



(d) Efficient Coding Explanation

**Figure 8.1** Four kinds of explanation.

In each case the explanandum, depicted at the top, is a biological phenomenon. (a) Aetiological Explanation: Smoking is said to *cause* the explanandum (lung disease). (b) Mechanistic Explanation: Ion channels opening and closing in response to changing membrane potential is said to *constitute* the explanandum (action potential). (c) Mathematical Explanation: The explanans is a mathematical fact (the Honeycomb Conjecture) and it can be thought of as *constraining* the path of evolution towards the optimal solution to the bees' storage problem. (d) Efficient Coding Explanation: The explanans is an abstract coding scheme which is said to be *implemented* by the actual retinal circuit.

To take an example which I discuss at greater length elsewhere (Chirimuuta 2017), it has been argued that the nervous system implements *hybrid* computation—a manner of processing information which alternates between analogue and digital codes (Sarpeshkar 1998). One property of hybrid computation is that it is energy efficient, using little power for each bit of information processed, in comparison with digital computation, while being less easily impacted by noise than purely analogue computation. Sarpeshkar argues that the implementation of hybrid computation explains how

biological brains can consume orders of magnitude less energy than man-made supercomputers, while being equivalent in computational capacity.

Here the explanandum is a particular behavior or feature of a neural system, namely the economy with which nervous tissue consumes energy. The explanans is a coding scheme, an abstractly characterized method of performing computations which has certain properties of its own, such as economical consumption of resources. There are mathematical frameworks, such as information theory, which tell us why the explanans has the property of interest. Physiological data are offered to provide evidence that the neural system *implements* the coding scheme. It is then argued that the reason why the neural system has the property of interest is that it is an implementation of the coding scheme theoretically shown to have this property. We then have an explanation of why the nervous tissue has the property in question.

This explanation is non-mechanistic because it does not proceed by decomposing the neural system and describing how the different component parts interact to give rise to the explanandum phenomenon. This idea that mechanistic explanations work by tracing the causal relationships between components of a tightly knit biological system is also encapsulated in the “models to mechanism mapping” (3M) criterion:

In successful explanatory models in cognitive and systems neuroscience (a) the variables in the model correspond to components, activities, properties, and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the perhaps mathematical dependencies posited among these variables in the model correspond to the perhaps quantifiable causal relations among the components of the target mechanism.

(Kaplan and Craver 2011: 611; cf. Kaplan 2011, 347)

The 3M criterion was introduced as part of an argument that all genuinely explanatory models in computational neuroscience are mechanistic ones. It is important to study efficient coding models because we find cases of explanation without 3M-style mapping (Chirimuuta 2014: 145). For example, with hybrid computation, we are not told how particular components of the coding scheme relate to a neural system, as unearthed through physiological and anatomical study.

One might object that implementation is itself a kind of mapping relationship, and so efficient coding explanations satisfy the 3M criterion for mechanistic explanation. However, this argument misses the point that the central feature of mechanistic explanation is the tracing of causal relationships between the components of the explanans—the presentation of a mechanistic description—and showing how this set of relationships is responsible for some of the causal properties of the explanandum phenomenon. In the case of efficient coding explanation, the explanans itself (not just the representation of it)⁶ is a mathematical object, namely, a coding scheme or algorithm; the explanans is not a set of entities and activities in a biological system. Moreover,

⁶ I say this because in the case of mechanistic explanation the mechanistic description may be presented as a mathematical equation, which is a representation of concrete entities and the causal processes occurring amongst them.

the relationship of implementation is not the constitutive one that is required for mechanistic explanation. We cannot say that the coding scheme “produces, maintains, or underlies” the neural phenomenon; instead, the neural system is just an instance of the coding scheme, realized in biological hardware.

Even if efficient coding explanations are non-mechanistic, one may still wonder if they are causal. Here things become a little complex. As has been noted elsewhere, when scientists present explanations of evolved systems which are subject to biological, physical, and mathematical laws, different kinds of explanations often rub shoulders and one can shift between causal and non-causal explanations with subtle changes in the specification of the explanandum (Andersen 2016; Chirimuuta 2017). For example, the explanation of *why honeycomb is hexagonally shaped* must cite both the causal biological facts that there is evolutionary pressure on honeybees to maximize storage volume and minimize building materials in making combs, as well as the mathematical argument that a hexagonal structure is the one which achieves this aim. However, the explanation of *why honeycomb is the best structure, given the bees’ needs* is “distinctively mathematical” (Lange 2013: 499–500).

In the case of hybrid computation, there is a causal (biological) explanation of why economy of computation is such an important factor in explaining nervous systems, whereas the explanation of why hybrid computation is optimal for biological brains is a non-causal one, based on principles of information theory (Chirimuuta 2017). So even if efficient coding explanations do not sit exclusively in the non-causal category, they do look “beyond causation” in a way that mechanistic explanations do not.

Before closing this section I would like to point out that all four kinds of explanation have the resources to answer what-if-things-had-been-different questions. In the case of mechanistic and aetiological explanation, we can conduct (real or hypothetical) experiments on the biological systems and observe how interventions on the explanans result in changes to the explanandum. While no one could intervene on the laws of mathematics, mathematical explanations do yield *counterpossible* information about how things would be different under these impossible scenarios (Baron et al. 2017). Efficient coding explanations address w-questions by telling us how things would be different under a range of either counterfactual or counterpossible scenarios. I will now present examples of efficient coding explanations in neuroscience, and then discuss actual and potential applications.

3. Lateral Inhibition and Explanations of Early Visual Responses

Retinal ganglion cells (RGCs) are the “output” neurons of the mammalian retina. It has long been observed that these neurons have a center-surround receptive field (RF) organization. For an ON-center RGC, when light falls in a certain small, circular area of the visual field, the neuron’s rate of firing will increase; and if light falls in the wider

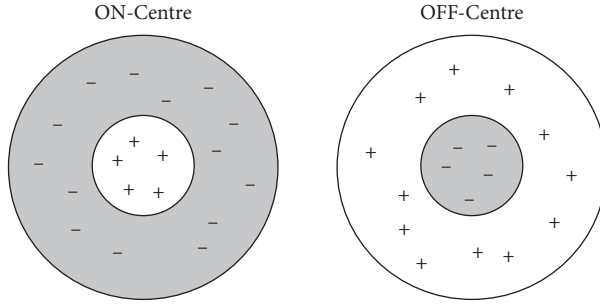


Figure 8.2 Receptive fields of retinal ganglion cells.

If light falls on the excitatory centre of an ON cell, firing rate will increase, whereas rate decreases if light falls on the inhibitory surrounding area. The polarity of responses is reverse for OFF cells.

area surrounding the center, then the firing rate will tend to decrease. OFF-center RGCs have the same concentric receptive field organization, but with opposite polarity (see Figure 8.2).

The Difference-of-Gaussian (DoG) function is commonly used to model the RF shape. For an ON-center cell, the first Gaussian function describes the response of the excitatory center, with A_1 (height of Gaussian) being the cell's maximum response and σ_1 (spread) describing the spatial extent of the center. The second Gaussian function, modeling the inhibitory surround, is subtracted from the first. The strength of inhibition is described by A_2 , and this takes a lower value than A_1 . σ_2 describes the spatial extent of the inhibitory surround, which takes a greater value than σ_1 . The DoG model is a two-dimensional, circularly symmetrical function in the x, y plane, centered at $(0,0)$:

$$F(x, y) = \frac{A_1}{2\pi\sigma_1^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_1^2}\right) - \frac{A_2}{2\pi\sigma_2^2} \exp\left(-\frac{x^2 + y^2}{2\sigma_2^2}\right) \quad (1)$$

In his discussion of the DoG function, David Kaplan argues that it is a phenomenological model with high predictive and descriptive value but lacking explanatory force. Explanations of the neurons' responses, it is argued, will be arrived at once we have modifications of the model which include mechanistic detail:

Transforming the DOG model... into an explanatory mechanistic model involves delimiting some of the components and some of the causal dependencies among components in the mechanism responsible for producing the observed structure of the receptive fields, along the lines indicated by 3M. One way to do this, for instance, would be to supplement the model with additional terms corresponding to various components in the retinal... circuit giving rise to the observed response properties of ganglion... neurons. (Kaplan 2011: 360)

Kaplan then references two neuroscientific articles on the retina which proceed in this direction. In contrast with this mechanistic perspective on the system, I will discuss a tradition of research which explains the neurons' response properties in terms of the

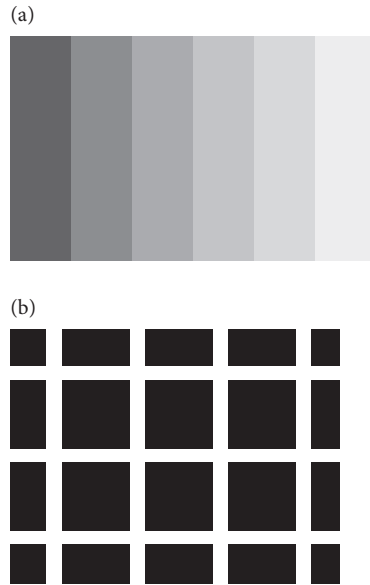


Figure 8.3 Visual illusions explained by lateral inhibition.

(a) Mach Bands: Within each of the broad vertical bands the grey level is uniform, yet we perceive a thin dark vertical strip near the border with a lighter band, and a thin lighter grey strip near the border with a darker band. (b) Hermann Grid: The dark spots at the intersections of the white crosses are illusory. In both cases the illusory patterns are attributed to the presence of inhibitory connections between retinal neurons (Ratliff 1961: 195).

(Source: Wikimedia commons)

information processing functions which they perform. This approach proceeds not by adding mechanistic detail to the DoG model but by interpreting it as implementing a particular coding strategy. We should think of the approach as addressing a very different kind of question from the one answered by mechanistic neuroscience—the question of *why* neural systems have the properties that are observed.⁷

The first step is to introduce the concept of *lateral inhibition*. Sensory neurons are said to exhibit lateral inhibition when excitation of one neuron brings about inhibition of the responses of its neighbors. The center-surround RFs of the retina are indicative of a circuit with lateral inhibition, since the suppressive areas of the RFs arise from the inhibitory inputs of nearby interneurons whose RFs are adjacent in the visual field. Lateral inhibition in the retina is the standard explanation of the visual illusions shown in Figure 8.3, and it is interesting to note that Ernst Mach posited that the Mach Band

⁷ This is a similar contrast to the famous ‘how?’ vs. ‘why?’ division in biology. As Barlow (1961b: 782) writes, Ratliff’s experiments on the crab’s eye “tell us a good deal about *what* the lateral inhibitory mechanism does and something about *how* it does it, but there remains a third question to ask. The fact that this mechanism has evolved independently in a wide variety of sensory relays suggests that it must have considerable survival value: *why* is this so?” Interestingly, this was published the same year as the institutionalization of the proximate/ultimate distinction by Mayr (1961).

illusion was caused by an antagonistic response arrangement in the visual system nearly a century before direct neural recordings were made.⁸

This sounds like the description of a mechanism and one might think that the explanation of Mach bands and the Hermann grid would look to be a just a mechanistic one. However, since the 1960s neuroscientists have offered at least three different non-mechanistic explanations for the presence of center-surround receptive fields and lateral inhibition in the early visual system. These non-mechanistic explanations all refer to the information processing task that has to be performed by the system, and they argue that lateral inhibition serves an important function in the service of this task.

3.1 *Edge detection*

All of the efficient coding explanations of lateral inhibition that I will discuss start with the idea that the early visual system must recode the input coming from the photoreceptors and suppress the signals which are not of high value to the downstream visual areas. One can think of the recoding by analogy with image processing routines which reduce the file size of a digital photograph. The data compression can either be “lossy” or “loss-less”. The first two proposals regarding lateral inhibition differ crucially in where they stand in the “lossiness” of the recoding. The edge detection hypothesis supposes that lateral inhibition serves to detect and/or enhance visual input that is most important to the downstream system—i.e., the edge structure in the visual scene—at the expense of passing on the rest of the input from the receptors. This is a lossy code because non-edge information is suppressed by lateral inhibition and cannot later be recovered by the downstream system.⁹

Two well-known proponents of the edge detection explanation of lateral inhibition are computer vision pioneers David Marr and Ellen Hildreth. Rather than beginning with neuroscientific findings, their approach to vision inquires “directly about the information processing problems inherent in the task of vision itself” (Marr and Hildreth 1980: 188; Marr 1982). As they see it, the task of the early visual system is to produce, from the raw photoreceptor input, a “primal sketch” of features such as edges, bars, and blobs. They show that one way to achieve this is by processing the input image with “Laplacian of Gaussian” filters, mathematical operators which find the areas of steepest illumination change—typically the edges in the image. Their filters are very similar to Difference of Gaussian functions discussed above, and are identical under certain parameter settings (Marr and Hildreth 1980: 207, 215–17). What kind of explanation of lateral inhibition is this? We are told that the function of the early visual circuits is to detect the edges that are present in the visual scene but are not represented sharply enough in the first encoding at the photoreceptor layer. This part of the story must appeal to causal processes—the system has the features that it does because it

⁸ See Weiskopf (2011). The effects of lateral inhibition have been observed in other perceptual modalities, like touch (von Békésy 1967: 41–5).

⁹ See Ratliff (1961: 183) and von Békésy (1967: 7). Barlow (1961a: 219) calls this the “password hypothesis”.

evolved or developed to perform a specific task. However, Marr and Hildreth also present a series of arguments and mathematical proofs to show that the image processing steps performed by their Laplacian of Gaussian operator is the optimal way to achieve the required representation of edges. This is a mathematical and non-causal explanation of why having neurons with the appropriate kind of lateral inhibition—those which implement the Marr–Hildreth operator—is the optimal way for the eye to achieve the desired task.

3.2 *Redundancy reduction*

The locus classicus for explanations of sensory physiology in terms of redundancy reduction is Horace Barlow's (1961a) paper, 'Possible Principles Underlying the Transformation of Sensory Messages'.¹⁰ Barlow draws on the influential article by Attneave (1954) which applies Claude Shannon's calculation of the redundancy of written English to the analysis of natural visual stimuli. Information theory provides the mathematical framework for thinking about neural signaling and redundancy. The basic idea is that "sensory relays" (of which retinal ganglion cells are an example) operate to recode information from inputs (ultimately—for RGCs—the photoreceptor layer), in such a way as to economize the consumption of resources (e.g., number of neurons needed, and number of action potentials fired on average). One way to economize is to reduce the redundancy of the code by eliminating signals which transmit information that is already known or expected by the receiver—see Figure 8.4. More generally, Barlow (1961a: 230) writes, "[t]he principle of recoding is to find what messages are expected on the basis of past experience and then to allot outputs with few impulses to these expected inputs, reserving the outputs with many impulses for the unusual or unexpected inputs."

We can see that the redundancy reducing code in Figure 8.4(b) is economical or efficient because it uses fewer action potentials to transmit the same amount of information as the first code (a). Since action potential generation is one of the major metabolic costs of the nervous system, it is reasonable to hypothesize that the nervous system, where possible, will operate in such a way as to minimize the number of spikes generated while maintaining the same rate of information transmission. This is how Barlow (1961a: 226) presents the hypothesis:

We may suppose that the [sensory] relay has a range of possible codes relating input to output: the [redundancy reduction] hypothesis says that, for a given class of input message, it will choose the code that requires the smallest average expenditure of impulses in the output. Or putting it briefly, it economizes impulses; but it is important to realize that it can only do this on the average; the commonly occurring inputs are allotted outputs with few impulses, but there may be infrequent inputs that require more impulses in the output than in the input.

¹⁰ Though as Barlow (1961a: 223) notes, the idea was prefigured in the writings of Karl Pearson, Kenneth Craik, Donald MacKay, and Ernst Mach.

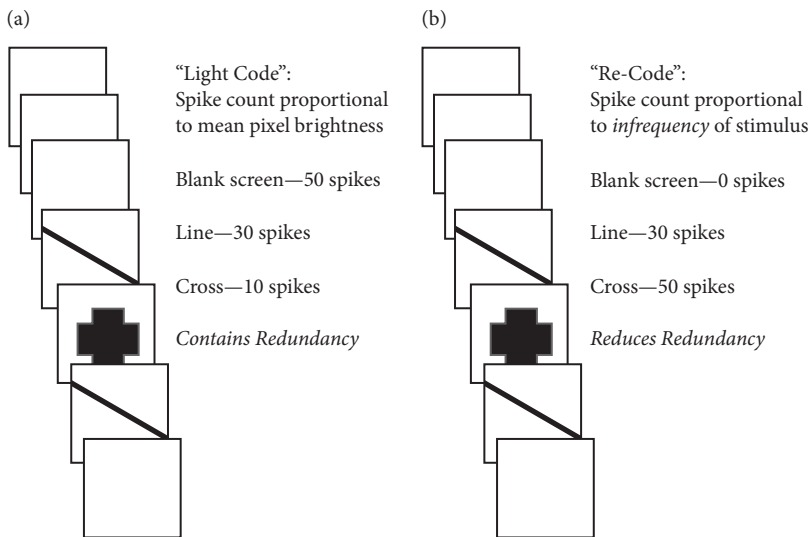


Figure 8.4 Re-coding to reduce redundancy.

(a) Light Code: Since neural response is proportional to mean pixel brightness, the blank screen will elicit the biggest response. But since the blank screen is frequent, and to be expected by the receiver of the signal, the spikes elicited by the blank screen are redundant. (b) Re-Code: Now the neural response is proportional to the *infrequency* of the stimulus. The blank screen is most frequent, so elicits no response; the cross is most infrequent, so causes the biggest response; and the response caused by the line is intermediate.

Note that this is a lossless code. The idea is not that the early visual system throws out, or makes unavailable, information that is there in the input concerning the most probable stimuli, but that it does not waste resources in signaling them to downstream receivers.

If we have reason to think that a neural system, like the retina, does indeed implement a redundancy reducing code, then we have an explanation for observed physiological properties, such as the receptive field structure of RGCs. Evidence for the implementation of a particular coding strategy typically comes in the form of physiological data about the system in question, anatomical findings about circuit structure, and a theoretical argument that the observed neural system can carry out the computation described by the coding scheme. Barlow (1961b: 782) himself argues that lateral inhibition is an effective means of attaining redundancy reduction via an example of photographic image processing.

We should now consider what kinds of explanation the redundancy reduction hypothesis provides. Again, there are both causal and non-causal dimensions. As apparent in Barlow's discussion of the different explanatory questions (see footnote 7), the redundancy reduction hypothesis is intended to explain what the evolutionary value of lateral inhibition is. Thus the resulting description of the information processing challenge that the retina faces, and the evolutionary pressure towards efficient coding, is a kind of (non-mechanistic) causal explanation. In a very abstract way, it

considers environmental conditions and selective pressures, and proposes that lateral inhibition is a result of these factors. For example, we are told that if there were no statistical regularities (spatial or temporal correlations) in natural visual stimuli (in the evolutionary environment of the animal) then the eye could not utilize a redundancy reduced code and we would not expect to see lateral inhibition.¹¹

Barlow's hypothesis also relies on the mathematical theory of information. The laws of information theory constrain the kinds of coding schemes that are efficient, given the actual environment and needs of the animal. In a non-causal sense, information theory 'makes a difference' to the kind of algorithm that the early visual system can implement. What if the laws of information theory were such that the system could reduce redundancy by making spike count proportional to the *frequency* of stimuli? Then you would not expect to have lateral inhibition because it would be efficient for the system to signal mean luminance. There is no way to intervene on laws of information theory, so this experiment is not even hypothetically possible. Yet Barlow's account gives us information about what would happen under such counterpossible scenarios.

For the purposes of this chapter, it need not matter whether this is a *good* explanation of retinal responses.¹² One theoretical reason for thinking that redundancy reduction is not the only "design principle" which can explain the mammalian retina and other early visual systems is the fact that redundancy reduction trades off against robustness to noise. This is easy to see if we take the example of a telegraph message being sent via an electric cable which experiences random fluctuations in the current or voltage. This noise will result in an error in the decoding of a proportion of the letters sent by the telegrapher. But because of the redundancy within written English (e.g., the regularity of a 'u' following a 'q'), up to a certain percentage of errors it is still quite easy to reconstruct the intended message. In other words, the code is robust to errors introduced due to noise. Since we know that neurons are noisy, this is bound to put constraints on the coding schemes employed by the nervous system.

3.3 *Predictive coding*

Unlike the others discussed so far, Srinivasan, Laughlin, and Dubs explicitly compare their explanatory hypothesis about the function of lateral inhibition with the alternative

¹¹ This fits the template of interventionist causal explanation. The redundancy reduction hypothesis tells us that statistical regularities in the visual environment make a difference to the coding schemes employed in the eye. One could perform a practically infeasible, but not modally impossible, experiment where one observes the evolution of creatures in an environment in which the only visual stimuli are random noise—i.e., no spatial or temporal correlations between visual inputs. We would not expect to see the development of lateral inhibition in early visual systems. In fact, Barlow's theory would probably predict the atrophy of the visual system, since under these conditions there is literally no visual information provided to the animal and so it cannot use this sensory modality to aid survival.

¹² For evidence that the retina does not always follow a redundancy reducing strategy because it fails to decorrelate the responses of neighboring RGCs, see Puchalla et al. (2005) but also Doi et al. (2012) and Borghuis et al. (2008). Barlow (2001) presents an extensive and deep criticism of his redundancy reduction argument.

proposals. Their claim is lateral inhibition as implements a *predictive code*,¹³ and that this account subsumes both the edge detection and redundancy reduction proposals (Srinivasan et al. 1982: 451). The idea is that the surround portion of the neuron's receptive field measures local mean luminance, giving a prediction of what the luminance will be in the center. If this prediction is accurate, then the luminance value at the center will be exactly cancelled out by the inhibitory input to the center, and the cell's firing will not increase. But if the central luminance value diverges from the prediction, then it will overcome the inhibition and a signal will be generated to say that something "surprising" is happening in the center. Unlike Barlow (1961a: 224), they also emphasize that lateral inhibition, understood in their way, has advantages for systems like the brain which have high intrinsic noise (Srinivasan et al. 1982: 427).

Srinivasan et al. (1982: 428) point out that the idea of predictive coding first came from television engineers in the 1950s. The predictive coding hypothesis has recently been employed by Sterling and Laughlin (2015: 249) in their comparison of early visual processing in mammals and flies. They write that, "predictive coding, an image compression algorithm invented by engineers almost 60 years ago to code TV signals efficiently, is implemented in animals by a basic sensory interaction". Once again, the idea is that we formulate an explanation of why the neural circuit has an observed feature by showing that it implements an algorithm known to be efficient—both in biological and artificial systems.

As in the previous two examples, there are both causal and non-causal features to this explanation. Sterling and Laughlin (2015: 249) place much emphasis on the tight energy budget of the central nervous system. This is a causal explanation of neural design, which tells us that if the energy budget were more ample, or if spikes cost fewer molecules of ATP, then we could expect different circuits. Alongside this reasoning, there is the mathematical argument that predictive coding is an efficient means to transmit visual information. This reasoning explains why a neural circuit for visual signal transmission, with a tight energy budget, would be constrained to implement predictive coding through lateral inhibition.

3.4 *Some observations*

Before moving on, I would like to say a few words about what we have learned from this case study. I have described the development of efficient coding explanations of one neural phenomenon, lateral inhibition, in order to make the case that this approach has been an active area of research, alongside the mechanistic one, since the very beginnings of physiological investigation of the visual system. In other words, as soon as neuroscientists were able to measure the effects of visual stimulation on

¹³ There has been much discussion in recent philosophy of the proposal that predictive coding provides a single unified framework for understanding mind and brain. See Hohwy (2013) and Clark (2016). Note that the proposal of Srinivasan et al. (1982) is much more modest in that it only extends to one specific circuit, and much more concrete in that it tells us exactly how the predictive code could be implemented by the circuit in question.

specific neurons in the early visual system, and plot their receptive fields, they began theorizing about the functions of those RFs and discussing abstract coding schemes which could be said to be implemented by the neural circuit. Researchers taking this approach have been very much in the mainstream of visual neuroscience.

The other point I would like to make here is that in each of the cases presented above, ideas about what the visual system was coding, and why, have been inspired quite directly by work outside of neuroscience: information theory and signal engineering, computer vision and television engineering. Do the origins of the efficient coding approach in engineering shape the practical applications of its findings? How are the reverse engineering of the brain and the forward engineering of brain-like machines connected?

4. Putting Efficient Coding Explanations to Use

4.1 *Scaling the data mountain*

Neuroscience does not suffer from a poverty of data. According to Hill (2015: 113), the rate of publication in neuroscience has grown from 30,000 articles per year in 1990 to 100,000 per year in 2013. What's missing is the means for neuroscientists to streamline and consolidate the deluge of results so that it is clear to each subfield what is known and what is not known.

At the beginning of their recently published book on the efficient coding approach to neural systems, Sterling and Laughlin (2015) are clear that they see their work as offering ways to digest the surfeit of data—or to switch to their metaphor, to climb the mountain of data. Their strategy is to articulate a small number of “organizing principles” that afford efficient coding explanation of diverse features of biological information processing in organisms spanning the chain of being, including bacteria, flies, and human brains.¹⁴ Many of these “design principles” come directly from engineering and information theory, while others are based on direct measurement of the cost of information processing in biological tissue. The basic idea is that by focusing on the information processing *function* of neural systems, scientists will be better able to discern the really important phenomena against the background of extraneous mechanistic detail.¹⁵

Interestingly, this motivation for the efficient coding approach was already stated by Barlow (1961a: 217):

A wing would be a most mystifying structure if one did not know that birds flew... [W]ithout understanding something of the principles of flight, a more detailed examination of the wing

¹⁴ They list ten such principles: “compute with chemistry; compute directly with analog primitives; combine analog and pulsatile processing; sparsify; send only what is needed; send at the lowest acceptable rate; minimize wire; make neural components irreducibly small; complicate; adapt, match, learn, and forget” (Sterling and Laughlin 2015: ii).

¹⁵ This sentiment is echoed by Marcus and Freeman (2015: xii), quoted at the start of section 4.3.

itself would probably be unrewarding. I think that we may be at an analogous point in our understanding of the sensory side of the central nervous system. We have got our first batch of facts from the anatomical, neurophysiological, and psychophysical study of sensation and perception, and now we need ideas about what operations are performed by the various structures we have examined. . . .

It seems to me vitally important to have in mind possible answers to this question when investigating these structures, for if one does not one will get lost in a mass of irrelevant detail and fail to make the crucial observations.

From our study of lateral inhibition we can already see how efficient coding explanations can be used to streamline and consolidate neuroscientific facts. As pointed out earlier, the eyes of mammals, crustaceans, and insects vary quite considerably in their anatomical and physiological details. By focusing on the *what?* and *how?* questions one could get lost in the mechanistic detail of each eye's neural circuit: the layout of the neurons, their dendritic arbors¹⁶ and activity patterns. In contrast, if one focuses on the question of *why* the neurons of a particular eye form an inhibitory network, and formulates an efficient coding explanation, the mechanistic details recede to the background and the similarities across mechanistically diverse systems become apparent.¹⁷ The key explanandum phenomenon is the kind of information processing that the inhibitory network affords, and since the explanans is an abstract coding scheme we need not worry too much about the details of biological implementation in each case (so long as a proposed implementation is not inconsistent with the known data).

This has echoes of the idea that explanation proceeds by showing that a set of seemingly unrelated phenomena can be unified with the same explanatory model or theory (Kitcher 1981). In fact, this remark by Hempel on explanation and unification is very much of a piece with Sterling and Laughlin's stated aims:

What scientific explanation, especially theoretical explanation, aims at is not [an] intuitive and highly subjective kind of understanding, but an objective kind of insight that is achieved by a systematic unification, by exhibiting the phenomena as manifestations of common, underlying structures and processes that conform to specific, testable, basic principles.

(Hempel 1966: 83, quoted by Kitcher 1981: 508)

I should note, however, that Sterling and Laughlin's declared inspiration is not twentieth-century philosophy of science but the unsurpassed subsumption of disparate data under unifying theory that was afforded by the theory of natural selection

¹⁶ As it happens, one ongoing project in retinal anatomy that has received much attention (and criticism) is Sebastian Seung's crowdsourcing challenge to get the complete wiring diagram (*connectome*) of the mouse retina. Much criticism has focused on the point that there is so much difference in the detailed anatomy even amongst individuals of the same species, that a dense reconstruction of the wiring cannot be practically or theoretically informative. But see Kim et al. (2014).

¹⁷ This point bears thinking about in relation to the argument of Weiskopf (2011) that lateral inhibition is a functional kind which is multiply realized in diverse systems—compound eyes like those of the fly and horseshoe crab, and the lens eyes of mammals. However, note that nothing in my argument turns on whether or not the multiple realization thesis is correct.

(Sterling and Laughlin 2015: xiv). Moreover, the explanatory sufficiency of efficient coding reasoning does not thereby stand and fall with the covering law and unificationist model of explanation. As I have been careful to point out, efficient coding explanations satisfy the requirement of answering w-questions, a condition which many critics of covering-law explanation subscribe to.

4.2 *Forward engineering*

Sterling and Laughlin's goal is to reverse engineer the brain. They do not discuss ways that the efficient coding approach could be applied beyond basic neuroscience, in neuro-inspired technologies and bio-engineering involving the brain. However, this is an increasingly active field of research and it is interesting to see how efficient coding explanations play a role in it.

More specifically, the concepts of efficient coding explanation—e.g., constraints, trade-offs, efficiency, redundancy, and optimization—come ultimately from engineering. While computational neuroscientists are taking a design stance to neurobiological systems and doing the reverse engineering, the principles that they formulate or discover (see footnote 14) will often apply equally to man-made systems and biological ones. This is necessarily the case when the principle in question is a result derived from information theory or any kind of mathematical or statistical argument. The trade-offs revealed by the mathematical analysis of information transmission can be thought of as design constraints that an information engineer ought to be conscious of, and knowledge of biological “solutions” frequently inspires better design. So even when trade-offs, such as the one between redundancy and robustness, cannot themselves be subject to intervention, knowledge of those trade-offs can have very direct practical application.

One of the spurs for studying the coding schemes which allow the brain to process information with much less power consumption than computers is the need to design more efficient artificial devices. Rahul Sarpeshkar, whose hybrid coding argument was discussed earlier, is an electronics engineer with a research focus on low-power biology-inspired computation. For example, his ideas have applications in the design of implantable medical electronics such as sensory-substitution devices (Sarpeshkar 2010). In the field of vision science we can note the influence running from engineering to neuroscience and back again. We saw in our case study of lateral inhibition, neuroscientists borrowed concepts from signal engineering and information theory in order to explain their observations. From the 1970s onwards there have been concerted efforts to design algorithms which will give computers or robots functioning vision. Though Marr (1982) famously argued that computer vision research was best off proceeding independently of visual neuroscience, bracketing questions about neural implementation, I think we should understand this as a warning against focusing on irrelevant mechanistic issues. Marr and Hildreth (1980) emphasize the comparison between their Laplacian of Gaussian filter and empirical findings in psychology and neuroscience about the workings of the early visual system, where these findings

concern the abstract coding schemes employed here rather than detailed anatomy or physiology.¹⁸

Another example is the use of the Gabor function to model the neurons in primary visual cortex (see Chirimuuta 2014: §5.2 and Chirimuuta forthcoming: §3). The introduction of the function, borrowed from mid-twentieth-century communications engineering, was justified by Daugman (1985) as the optimal solution to the joint problem of decoding both spatial location and spatial frequency (width of edge) information. John Daugman is a computer scientist who has sought to design better image recognition algorithms on the basis of his study of visual cortex.

Furthermore, the engineering approach can also be applied to the manipulation of the brain itself, not just in the building of artificial devices. Neuro-engineering is a fast-growing field of activity involving the development of *brain-computer interfaces* (BCIs) which read off and decode neural activity in order to control external devices such as computers and robotic limbs, or to channel information directly into the brain. In order for such technologies to be effective, the brain's activity must be understood in abstract enough terms to allow for translation to and from digital computers. That is, the “neural code”—the information conveyed by particular patterns of activity—must be deciphered and manipulated in a way that is independent of the specific biological implementation (Chirimuuta 2013). This is why abstraction from mechanistic details, and recourse to rarefied mathematical descriptions of signals, is particularly useful here. Yet in order to build an effective BCI, a brilliant decoding algorithm is not enough. One also needs an electrode implant in the cortex which has long-term stability and does not quickly lead to degeneration of the neural tissue in which it is embedded. Of course this requires precise anatomical knowledge of the cortical layers, knowledge of the biochemical environment, and of neural cell death cascades—in other words, a detailed mechanistic understanding of the brain. This is a field of endeavor in which mechanistic and efficient coding knowledge are both integral to its success.

4.3 *Defining neural computation*

It is uncontroversial, amongst neuroscientists, to say that the brain computes (Koch 1998: 1). And it is by now well established that the brain does not compute in the same way that a general purpose digital computer does, or in the fashion of any known analogue machine. I concur with Piccinini and Bahar (2013: 476) that neural computation is *sui generis*. The tricky thing is then to put some useful definitions in place which will help clarify what is or should be meant by neural computation, and there is not yet a consensus emerging from the discipline of theoretical neuroscience. As Marcus and Freeman (2015: xii) write, “we have yet to discover many of the organizing

¹⁸ Note also that computer vision algorithms which employ lateral inhibition—e.g., by using the DoG function—are quite commonly used. See Klette (2014: 75–6), Moini (2000: 18–19), and Lyon (2014) on the invention of the optical mouse.

principles that govern all that complexity. We don't know, for example, if the brain uses anything as systematic as, say, the widespread ASCII encoding scheme that computers use for encoding words. And we are shaky on fundamentals like how the brain stores memories and sequences events over time."

Piccinini and Bahar (2013: 477–9) assert that computation is a kind of "mechanistic process", and thus that the empirical study of neural mechanisms, and the search for mechanistic explanations of the brain and psychological states, will eventually lead to an understanding of neural computation. I believe that this approach is misguided. As we saw in the case study of lateral inhibition, any restricted focus on the mechanistic details giving rise to inhibitory effects would not be illuminating as to the computational properties of the circuit. For one thing, the search for mechanistic explanations does not draw from the theoretical frameworks in engineering and mathematics which can be used to characterize computational systems.¹⁹ For another, the mechanistic perspective obscures the interesting commonalities amongst biophysically very different systems. It was only by taking the efficient coding perspective, and asking in abstract terms what function the circuit performs, and why, that hypotheses could be formed about what coding scheme is implemented in these systems.

In order to make progress towards a definition and theory of neural computation, general coding schemes and unifying principles are far more valuable than a disunified collection of data concerning mechanisms in the brains of different animals. This requires that scientists work with a "level of description" which is abstracted from that of mechanistic implementation (cf. Marr 1982; Carandini 2012), and is assumed in the efficient coding tradition. One idea along these lines which has recently been attracting attention is that of *canonical neural computations* (Carandini and Heeger 2012). These are computational operations which are frequently used to model small circuits and are found to reoccur in different species and brain regions. The DoG model of lateral inhibition would be an example, and they are commonly invoked in efficient coding explanations. Carandini and Heeger's proposal is to identify a handful of such computations which might be thought of as the building blocks for more complex neural computations. If the project is successful, the result would be a clearly articulated theory of neural computation.

5. Conclusion

In this chapter I have charted the development of efficient coding explanations of a well-known neural phenomenon, and discussed practical applications of these and other models and explanations. I have been somewhat diffident about the causal/non-causal distinction because in practice these aspects of efficient coding explanation are integrated and complementary to one another. What is more significant is the difference between efficient coding and mechanistic explanation, since each approach

¹⁹ But see Koch (1998) for a hybrid computational–mechanistic approach.

reveals and obscures different aspects of a neural system. For example, efficient coding models tend to mask the bio-chemical intricacy of the brain's 'circuits', treating them more like arrays of electronic switches. As a result, such models do not play a role in the development of pharmaceuticals to alleviate organic diseases affecting brain cells; they do make a difference, however, in the design of prosthetic systems which aim to replace lost neural tissue. More generally, they have an important place in tasks where 'big picture' ideas about the system's function are needed.

Throughout this chapter I have emphasized the extent to which the efficient coding framework draws from the theories and concepts of communication engineering. I would like to finish with the caveat that this analogical approach to understanding the brain brings with it its own limitations. Both neuroscientists and philosophers of neuroscience should be aware of the ways in which the analogy between the brain and a man-made computer or signaling system can break down. As Barlow (2001: 244) puts it, "[i]n neuroscience one must be cautious about using Shannon's formulation of the role of statistical regularities, because the brain uses information in different ways from those common in communication engineering." The challenge is to find out exactly *how* the brain uses information, and what "information" is in the context of neuroscience rather than engineering. The efficient coding approach is just a starting point.

Acknowledgments

I would very much like to thank Peter Sterling and the editors of the volume for many thoughtful comments and their help in improving this chapter.

References

- Andersen, H. (2016), 'Complements, Not Competitors: Causal and Mathematical Explanations', *British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axw023.
- Attneave, F. (1954), 'Some Informational Aspects of Visual Perception', *Psychological Review* 61: 183–93.
- Barlow, H. B. (1961a), 'Possible Principles Underlying the Transformation of Sensory Messages', in W. A. Rosenblith (ed.), *Sensory Communication* (Cambridge, MA: MIT Press), 217–34.
- Barlow, H. B. (1961b), 'Three Points about Lateral Inhibition', in W. A. Rosenblith (ed.), *Sensory Communication* (Cambridge, MA: MIT Press), 782–6.
- Barlow, H. (2001), 'Redundancy Reduction Revisited', *Network* 12: 241–53.
- Baron, S., Colyvan, M., and Ripley, D. (2017), 'How Mathematics Can Make a Difference', *Philosophers' Imprint*.
- Bokulich, A. (2011), 'How Scientific Models Can Explain', *Synthese* 180: 33–45.
- Borghuis, B. G., Ratliff, C. P., Smith, R. G., Sterling, P., and Balasubramanian, V. (2008), 'Design of a Neuronal Array', *Journal of Neuroscience* 28: 3178–89.
- Carandini, M. (2012), 'From Circuits to Behavior: A Bridge too Far?', *Nature* 15: 507–9.

- Carandini, M. and Heeger, D. J. (2012), 'Normalization as a Canonical Neural Computation', *Nature Reviews Neuroscience* 13: 51–62.
- Chirimuuta, M. (2013), 'Extending, Changing, and Explaining the Brain', *Biology & Philosophy* 28: 613–38.
- Chirimuuta, M. (2014), 'Minimal Models and Canonical Neural Computations: The Distinctness of Computational Explanation in Neuroscience', *Synthese* 191: 127–53.
- Chirimuuta, M. (2017), 'Explanation in Computational Neuroscience: Causal and Non-Causal', *British Journal of the Philosophy of Science*. DOI: 10.1093/bjps/axw034.
- Clark, A. (2016), *Surfing Uncertainty: Prediction, Action, and the Embodied Mind* (Oxford: Oxford University Press).
- Daugman, J. G. (1985), 'Uncertainty Relation for Resolution in Space, Spatial Frequency, and Orientation Optimized by Two-Dimensional Visual Cortical Filters', *Journal of the Optical Society of America. A: Optics and Image Science* 2: 1160–9.
- Doi, E., Gauthier, J. L., Field, G. D., Shlens, J., Sher, A., Greschner, M., Machado, T. A., Jepsen, L. H., Mathieson, K., Gunning, D. E., Litke, A. M., Paninski, L., Chichilnisky, E. J., and Simoncelli, E. P. (2012), 'Efficient Coding of Spatial Information in the Primate Retina', *Journal of Neuroscience* 32: 16256–64.
- Hempel, C. G. (1966), *Philosophy of Natural Science* (Englewood Cliffs, NJ: Prentice-Hall).
- Hill, S. (2015), 'Whole Brain Simulation', in G. Marcus and J. Freeman (eds.), *The Future of the Brain* (Princeton, NJ: Princeton University Press), 111–24.
- Hohwy, J. (2013), *The Predictive Mind* (Oxford: Oxford University Press).
- Husbands, P. and Holland, O. (2008), 'The Ratio Club: A Hub of British Cybernetics', in P. Husbands, O. Holland, and M. Wheeler (eds.), *The Mechanical Mind in History* (Cambridge, MA: MIT Press), 91–148.
- Kaplan, D. M. (2011), 'Explanation and Description in Computational Neuroscience', *Synthese* 183: 339–73.
- Kaplan, D. M. and Craver, C. F. (2011), 'The Explanatory Force of Dynamical and Mathematical Models in Neuroscience: A Mechanistic Perspective', *Philosophy of Science* 78: 601–27.
- Kim, J. S., Greene, M. J., Zlateski, A., Lee, K., Richardson, M., Turaga, S. C., Purcaro, M., Balkam, M., Robinson, A., Behabadi, B. F., Campos, M., Denk, W., Seung, H. S., and the EyeWriters (2014), 'Space-Time Wiring Specificity Supports Direction Selectivity in the Retina', *Nature* 509: 331–6.
- Kitcher, P. (1981), 'Explanatory Unification', *Philosophy of Science* 48: 507–31.
- Klette, R. (2014), *Concise Computer Vision: An Introduction into Theory and Algorithms* (Dordrecht: Springer).
- Koch, C. (1998), *Biophysics of Computation: Information Processing in Single Neurons* (New York: Oxford University Press).
- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Lyon, R. F. (2014), 'The Optical Mouse: Early Biomimetic Embedded Vision', in B. Kisačanin and M. Gelautz (eds.), *Advances in Embedded Computer Vision* (Dordrecht: Springer), 3–22.
- Marcus, G. and Freeman, J. (2015), 'Preface', in *The Future of the Brain* (Princeton, NJ: Princeton University Press), xi–xiii.
- Marr, D. (1982), *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information* (San Francisco: W. H. Freeman).

- Marr, D. and Hildreth, E. (1980), 'Theory of Edge Detection', *Proceedings of the Royal Society of London. B: Biological Sciences* 207: 187–218.
- Mayr, E. (1961), 'Cause and Effect in Biology', *Science* 134: 1501–6.
- Moini, A. (2000), *Vision Chips* (Dordrecht: Kluwer).
- Piccinini, G. and Bahar, S. (2013), 'Neural Computation and the Computational Theory of Cognition', *Cognitive Science* 34: 453–88.
- Puchalla, J., Schneidman, E., Harris, R., and Berry, M. J. (2005), 'Redundancy in the Population Code of the Retina', *Neuron* 46: 493–504.
- Ratcliff, F. (1961), 'Inhibitory Interaction and the Detection and Enhancement of Contours', in W. A. Rosenblith (ed.), *Sensory Communication* (Cambridge, MA: MIT Press), 183–203.
- Saatsi, J. and Pexton, M. (2013), 'Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations', *Philosophy of Science* 80: 613–24.
- Sarpeshkar, R. (1998), 'Analog versus Digital: Extrapolating from Electronics to Neurobiology', *Neural Computation* 10: 1601–38.
- Sarpeshkar, R. (2010), *Ultra Low Power Bioelectronics* (Cambridge: Cambridge University Press).
- Sprevak, M. (2012), 'Three Challenges to Chalmers on Computational Implementation', *Journal of Cognitive Science* 13: 107–43.
- Srinivasan, M., Laughlin, S., and Dubs, A. (1982), 'Predictive Coding: A Fresh View of Inhibition in the Retina', *Proceedings of the Royal Society of London. B: Biological Sciences* 216: 427–59.
- Sterling, P. and Laughlin, S. B. (2015), *Principles of Neural Design* (Cambridge, MA: MIT Press).
- von Békésy, G. (1967), *Sensory Inhibition* (Princeton, NJ: Princeton University Press).
- Weiskopf, D. A. (2011), 'The Functional Unity of Special Science Kinds', *British Journal for Philosophy of Science* 62: 233–58.
- Woodward, J. F. (2003), *Making Things Happen* (New York: Oxford University Press).

9

Symmetries and Explanatory Dependencies in Physics

Steven French and Juha Saatsi

1. Introduction

In this chapter we will investigate explanations that turn on symmetries in physics. What kinds of explanations can symmetries provide? How do symmetries function as an explanans? What philosophical account of explanation can naturally capture commonplace symmetry-based explanations in physics? In the face of the importance and prevalence of such explanations and symmetry-based reasoning in physics, it is striking how little has been written about these issues.¹ It is high time to start examining these hitherto largely ignored questions.

In this chapter we will argue that various symmetry explanations can be naturally captured in terms of a *counterfactual-dependence* account in the spirit of Woodward (2003), liberalized from its causal trappings. From the perspective of this account symmetries can function in explanatory arguments by playing a role (roughly) comparable to a contingent initial or boundary condition in causal explanations: a symmetry fact (in conjunction with an appropriate connection between that fact and the explanandum) can contribute to provision of what-if-things-had-been-different information, showing how an explanandum *depends on* the symmetry. That is, symmetries can explain by providing modal information about an explanatory dependence, by showing how the explanandum would have been different, had the facts about the symmetry been different.

Explanatory dependencies of this sort need not be causal. Although the counterfactual-dependence view of explanation is best developed in connection with causal dependence, in recent years this view has been extended to various kinds of non-causal dependencies (e.g., Jansson and Saatsi forthcoming; Reutlinger 2016; Saatsi forthcoming; Saatsi and Pexton 2013). Our discussion of symmetry explanations is more grist to this

¹ Lange's work on symmetry principles and conservation laws is a notable exception (e.g., Lange 2007, 2012).

mill: many (but not all) symmetry explanations are naturally construed as being non-causal, as we will see. But even if symmetry is not a cause of an explanandum, we may nevertheless be able to regard the explanandum as something that depends in an explanatory way on the symmetry in question. Or so we will argue.

There are alternative accounts of explanation that compete with our counterfactual-dependence perspective, especially in the context of non-causal explanations that are highly abstract or mathematical (Pincock 2007, 2014; Lange 2013; cf. Jansson and Saatsi forthcoming for discussion). One alternative is to operate in the unificationist tradition of Friedman (1974) and Kitcher (1981, 1989). However, this faces well-known problems, not the least of which concerns the heterogeneity of unificatory practices (see e.g., Redhead 1984). In the case of symmetries in physics in particular, although their unificatory force is obviously connected to their heuristic role (as evidenced through the construction of the so-called Standard Model of particle physics) it is unclear how to cash out the unificatory force beyond that role. Of more current interest is a new approach to non-causal explanations developed by Lange (2007, 2012, 2013), who puts the explanatory weight on the independence of the explanandum from particular laws of nature. Interestingly, Lange has also applied this approach to some central issues concerning symmetry explanations. We will discuss Lange's views insofar as it runs contrary to our counterfactual-dependence account, but we will not attempt a broader assessment of these alternative viewpoints. We shall mainly endeavour to show that a counterfactual-dependence account can naturally deal with various symmetry-based explanations, thereby further supporting the now popular idea that explanations—causal and non-causal alike—provide information about worldly dependence relations that show what is responsible for the explanandum at stake. We will also discuss the extent to which this analysis of symmetry explanations requires us to relinquish the notion that all explanatory dependencies in science are causal (cf. Skow 2014).

The first order of business is to introduce the key notion, symmetry, and its connection to explanation (section 2). The rest of the chapter is divided between issues concerning the two basic kinds of symmetries found in science: discrete (section 3) and continuous (section 4).

2. Symmetry and Explanation: A Toy Example

What is *symmetry*, then? In very informal and general terms, the notion of symmetry involves *sameness* (or equivalence) in one respect of *X*, in relation to a *change* (or transformation) in another respect of *X*. What 'sameness in relation to change' exactly consists in is determined by the nature of *X*, the kind of transformation at stake, and in what respect it stays the same in relation to that transformation. Most familiar examples involve geometrical figures, spatial transformations (e.g., rotations), and the sameness of the figure (e.g., with respect to its shape) under those transformations. For instance, an equilateral triangle is thus symmetrical with respect to 120 degree turns. It is also

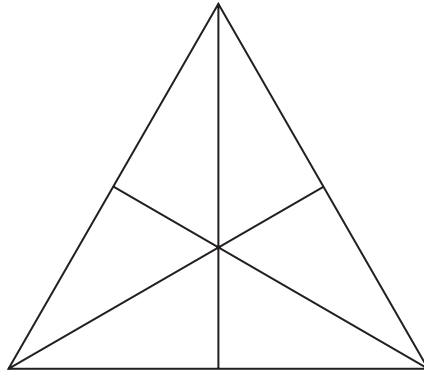


Figure 9.1 A symmetrical triangle.

symmetrical in relation to a transformation that reflects or flips the figure with respect to one of the three axes of symmetry (Figure 9.1).

More interesting objects of symmetry can involve things like laws of nature (or their mathematical expressions), which can retain their content (or form) under transformations of frames of reference (or coordinate systems). Regardless of the subject matter, symmetry can usually be made precise via the mathematical terms of group theory, where it is naturally defined as *invariance under a specified group of transformations*. The group theoretic framework makes precise the intuitive notion of ‘sameness in relation to change’ by showing how a symmetry group partitions the object of symmetry into equivalence classes, the elements of which are related to one another by symmetry transformations.²

With this notion of symmetry in mind, let’s look at a simple toy example of a symmetry, and a related explanation. Consider a balance (a see-saw, say), in a state of equilibrium (Figure 9.2). Assume the balance remains in the state of equilibrium when particular forces are applied on its two arms. Why does the balance remain in balance? How do we explain this? The standard answer is to appeal to the (bilateral) symmetry of the situation: there is an appropriate equivalence between the forces on the two

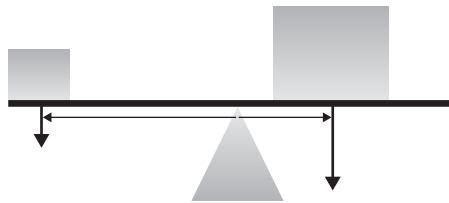


Figure 9.2 Balance in equilibrium.

² For details, see e.g., Olver (1995).

arms, so that the torque applied from each side to the pivot point is equal—namely, the net torque vanishes. Given this equivalence there are no grounds for the balance to move and hence it remains in equilibrium. Brading and Castellani (2003) call this a ‘symmetry argument’, and note that the lack of grounds can be understood as an application of the Principle of Sufficient Reason. Our interest lies in, first, the explanatory nature of the argument and second, and more importantly, in the role of symmetry as part of the explanans.

Let’s see how the symmetry argument could be accommodated in the counterfactual-dependence framework, which has at its core the idea that an explanation shows how the explanandum depends on the explanans. Can we find in the case of the balance an explanatory (asymmetric) dependence, associated with counterfactual information that answers what-if-things-had-been-different questions?

The answer is yes: the toy example fits the counterfactual-dependence account of causal explanation. The relevant physics is exceedingly simple, of course. The balance stays in a state of equilibrium if and only if the net torque on the pivot point is zero. This law-like connection between the (non-)equilibrium state of the balance and the forces involved obviously allows us to run the argument in both ways. On the one hand, from vanishing net torque we can deduce the state of equilibrium (assuming the balance was initially at rest). On the other hand, we can also deduce from a state of equilibrium the vanishing net torque. There is no asymmetry inherent in the law we employ in the explanation. (An attempt to capture the explanatory symmetry argument in the DN-model thus immediately runs into familiar problems regarding explanatory asymmetry.)

Nevertheless, intuitively there is an obvious explanatory asymmetry to be found: we can change the net torque (by intervening on the forces involved) so as to thereby change the (non-)equilibrium state of the balance, but not the other way around. That is, we cannot change the net torque through somehow acting on the (non-)equilibrium state of the balance, without intervening on the forces involved. That is why the vanishing net torque is not explained by the equilibrium state of the balance; it is only explained in terms of the forces that ‘sum up’ to zero. The counterfactual-dependence account of explanation, as developed by Woodward (2003), capitalizes on this explanatory asymmetry. In this case the counterfactual dependence involved has a natural interventionist-causal interpretation, of course. The explanation provides (high-level) information about the causes acting on the balance, and what would happen (*vis-à-vis* equilibrium) if the forces were different in the relevant ways.

What role does *symmetry* play in the explanation then? Although we are dealing with a causal explanation, there is clearly a sense in which the explanandum depends on a symmetry exhibited by the system. Since *any* non-zero net torque would move the balance to a non-equilibrium state, we can take as the relevant explanans a high-level feature of the system that abstracts away from lower-level information regarding the specific forces applied: all that matters for the explanation is whether or not there is a

bilaterally equivalent, *symmetrical* distribution of forces. There is thus a natural sense in which the equilibrium depends on symmetry. Now of course, in this case, there are forces involved, so what we have is a symmetry of such causal factors. Nevertheless, we will next argue that this example of symmetry explanation is not that different from other examples of symmetry based explanations where the existence of such fundamental causal factors is either questionable, at best, or entirely lacking.

3. Discrete Symmetries

The bilateral symmetry in the toy example above is an example of *discrete* symmetry. These are symmetries represented by groups involving discrete sets of elements (where these elements are typically enumerated by the positive integers). They frequently arise within physics, and include the well-known examples of Permutation Invariance and Charge-Parity-Time symmetry.

Let's begin with Permutation Invariance.³ To get an idea of what it involves, consider the standard example of two balls distributed over two boxes. Classically, we obtain four possible arrangements, but in quantum mechanics only three arise: both balls in the left hand box (say), both in the right hand box, or one ball in each. The crucial point is that a permutation of balls between the boxes is not counted as giving rise to a new arrangement, and it is upon this exemplification of Permutation Invariance that all of quantum statistics rests. In most textbooks on the subject this is taken to come in just two forms. Bose-Einstein statistics, which—in terms of our simple example—allows for both balls (or particles) to be in the same box (or state), applies to photons, for example. The alternative, Fermi-Dirac statistics, which applies to electrons, for example, prohibits two particles from occupying the same state. These two possibilities are encoded in what is generally taken to be a fundamental symmetry of quantum mechanics, captured by the 'Symmetrization Postulate', which says that the relevant wave or state function must be either symmetric—corresponding to Bose-Einstein statistics—or anti-symmetric—generating the Fermi-Dirac form. However, as is well-known, the mathematics of group theory allows for other possibilities, including the statistics of so-called 'paraparticles'.⁴ These further possibilities are encoded in a broader principle, known as Permutation Invariance, which, when applied to a particular system, dictates that the relevant Hamiltonian of the system must commute with the group theoretic particle permutation operator (French and Rickles 2003; French and Krause 2006).⁵ Although parastatistics do not appear in nature (as far as we

³ See French and Rickles (2003), and French and Krause (2006), for details.

⁴ 'Infinite' statistics are also allowed (Greenberg 1990) and in spaces of less than three dimensions one obtains 'braid' statistics and anyons.

⁵ Permutation Invariance thereby divides Hilbert space up into superselection sectors corresponding to the possible types of permutation symmetry associated with the different kinds of particles (bosons, fermions, para-bosons, para-fermions, and so on).

know)⁶ Permutation Invariance is generally regarded as the more fundamental symmetry principle (Messiah and Greenberg 1964).⁷

Now, consider the following as an example of the role of Permutation Invariance in an explanation. Those stars that develop into red giants but have masses less than four times that of the sun (which thus includes the sun itself) will in due course undergo a collapse, until they form a so-called 'white dwarf'. (White dwarves' average diameter is of the order of the Earth's diameter, and they have correspondingly massive density.) The explanation of the collapse has to do with the fact that such stars do not have sufficient energy to initiate the fusion of carbon (their hydrogen having been used up) and thus the balance between the gravitational attraction and the outward thermal pressure is disturbed, in favour of the former. However, there is a further phenomenon that demands explanation: why, at a certain point, does this collapse halt? The answer, given in the physics textbooks, is that this has to do with 'electron degeneracy', understood in this case as the result of the application to stellar statistical physics of Pauli's Exclusion Principle (PEP). The central idea is that according to PEP, no two electrons can be in the same state, and hence as the star contracts, all the lower energy levels come to be filled, so the electrons are forced to occupy higher and higher levels, which creates an 'effective pressure' that eventually balances the gravitational attraction.

The explanation of the halting of the white dwarf collapse thus critically turns on PEP. We regard it as a symmetry-based explanation since PEP, furthermore, drops out of the Symmetrization Postulate, which, we recall, requires the wave functions of all known types of particle to be either symmetric, yielding bosons, or anti-symmetric, corresponding to fermions, which behave according to Fermi-Dirac statistics. It is the latter anti-symmetry that gives rise to PEP. This distinction corresponds to perhaps the most fundamental natural kind distinction there is as fermions make up what we might call the 'material' particles, whereas bosons are the 'force carriers'.⁸

It has been suggested that this represents an example of a non-causal explanation of a physical phenomenon, given that Pauli's Principle puts a global constraint on possible states of the system.⁹ How should this explanation be understood? A number of philosophers have fretted over this question. Lewis, for example, talks of PEP as representing 'negative information' about causation:

⁶ Although it was suggested in the mid-1960s that quarks might be paraparticles of a certain statistical type, this was subsequently abandoned in favour of a description in terms of the property that became known as 'colour', leading to the development of quantum chromodynamics (French 1995).

⁷ It also grounds the well-known discussions of particle indistinguishability in quantum physics; see French and Krause (2006).

⁸ But as we also noted, the restriction to only symmetric and anti-symmetric wave functions is in fact a contingent feature of the world and other symmetry types are theoretically possible, corresponding to paraparticle statistics, as permitted by the broader requirement of Permutation Invariance.

⁹ Interestingly, physicists never call Pauli's Principle a 'law'. If considered as such, PEP is a *law of co-existence*, as opposed to a *law of succession*. The former restrict positions in the state-space, while the latter restrict trajectories in (through) the state-space. (See van Fraassen 1991: 29.) It is also a *global* constraint that concerns the universe as a whole, not some subsystem of it.

A star has been collapsing, but the collapse stops. Why? Because it's gone as far as it can go. Any more collapsed state would violate the Pauli Exclusion Principle. It's not that anything caused it to stop—there was no countervailing pressure, or anything like that. There was nothing to keep it out of a more collapsed state. Rather, there just was no such state for it to get into. The state-space of physical possibilities gave out. . . . [I]nformation about the causal history of the stopping has been provided, but it was information of an unexpectedly negative sort. It was the information that the stopping had no causes at all, except for all the causes of the collapse which were a precondition of the stopping. Negative information is still information. (Lewis 1986: 222–3)

Attempting to shoehorn this into the causal framework by suggesting that the lack of causal information is still indicative of causal relevance, might strike many as a desperate manoeuvre. Skow (2014), however, has recently argued that it can be brought into the causal framework, insisting, first, that it is not the case that the stopping had no causes at all and second, that there are in fact states for the electrons to 'get into'.

With regard to the first point, Skow notes that many physics textbooks standardly refer to the 'pressure' of a degenerate electron gas in this and other cases. He insists that there is, therefore, a sense in which we can attribute a countervailing pressure to the gravitational attraction, so that the explanation can be regarded as causal. It is important to note, as Skow himself does, that the so-called 'pressure' in this case is very different from that ascribed to a gas, say, since it is not due to any underlying electrostatic force, or indeed any force at all. Indeed, in the years following the establishment of PEP physics struggled to disentangle itself from the understanding of it in terms of 'exclusion forces' and the like (Carson 1996). Thus, one might be inclined to argue that the use of the term 'pressure' here is no more than a *façon de parler*, or a pedagogic device, and that in terms of our standard conception of pressure as grounded in certain causal features relating to the relevant forces involved (typically electromagnetic), there is simply no such thing as 'degeneracy pressure'.

Skow rejects such a move, insisting that terms in quantum statistical physics, such as 'pressure' and, indeed, 'temperature', have escaped their thermodynamic origins and must be conceived of in more abstract terms than as resulting from the force-based interactions of particles or as identical to mean molecular kinetic energy, respectively (2014: 458–9). Rather, according to Skow these terms should be regarded as dispositional: as the disposition of a system to transfer energy or 'volume', respectively, to another body. Thus, something other than repulsive forces between constituents—such as the consequences of PEP, for example—can contribute to the pressure of a system, rendering the 'degeneracy pressure' explanation causal, after all.¹⁰

¹⁰ With regard to Skow's second point, concerning Lewis's claim that the collapse stops because there is no state for the star as a whole to get into, Skow insists that this claim is also false (2014: 459–60). As he points out, what PEP excludes are states of the star in which more than one electron is in the same quantum state. However, he argues, since there are always infinitely many states available, the electrons never run out of states to get into (because there are always some empty ones available, albeit of high energy), no matter how small the star is. Hence the fact that the star stops collapsing at a certain size has nothing to do with the lack of available states for the electrons to occupy. According to Skow, "no matter how small the star's

Now we might just pause at this point and wonder whether ‘pressure’, characterized in such abstract terms, can be understood as appropriately causal. After all, in the case of the white dwarf star, this ‘transfer of volume’ still does not proceed via any of the known forces and it is unclear how to understand this notion in causal terms. Nevertheless, we shall be charitable and set these issues specific to statistical physics to one side as we believe there are reasons for thinking that non-statistical explanations essentially involving PEP clearly go beyond the causal framework.

Consider, for example, the explanation of chemical bonding. In 1927 Heitler and London explained the bonding in a homonuclear molecule such as H_2 by explicitly invoking PEP. It had become evident that the attraction between two hydrogen atoms could not be accounted for in terms of Coulomb forces; the key, as Heitler realized, lay with the so-called exchange integral, previously introduced by Heisenberg, which was something purely quantum mechanical, with no classical analogue (Gavroglu 1995: 45). Heitler and London proceeded from the fundamental basis that the electrons were indistinguishable and hence the usual way of labelling them when writing out the relevant wave function had to be rethought.¹¹ It then followed that the electronic wave function of the two-atom system had to be written in either symmetric or anti-symmetric form, according to the Symmetrization Postulate. With the electron spins incorporated, PEP dictates that the anti-symmetric form be chosen, with spins anti-parallel. This corresponds to the state of lower energy and attraction is thus understood on the basis of energy minimization. Thus, by deploying the Exclusion Principle chemical valence and saturation could be understood and the ‘problem of chemistry’ solved, or as Heitler put it, ‘Now we can eat chemistry with a spoon!’

This forms the basis of valence bond theory, further developed by Pauling and others, and which is now regarded as complementary to molecular orbital theory. Unlike the former, the latter does not assign electrons to distinct bonds between atoms and approximates their positions via Hartree-Fock or ‘Density Function’ techniques.

radius, the electrons never run out of states because there are infinitely many of them” (2014: 460). Thus, the cessation of the star’s collapse is “not because a state with a smaller radius is physically impossible, but because the star has reached the radius at which the outward-directed pressure in the star exactly balances the inward-directed gravitational forces. This is a paradigmatically causal explanation” (2014: 460). However, we think it is odd to insist that the radius of the star can be disassociated from the availability and occupation of electron states, since it is the latter that determine the former: the higher the energy state, or, putting it somewhat crudely, the further away the energy level is, the bigger the star. Skow is right in that the collapse stops when the star reaches a radius at which the degeneracy ‘pressure’ balances the gravitational attraction, but given that attraction (i.e., given the mass of the star) PEP ensures that it is impossible for the star to achieve a smaller radius, without a reduction in the number of particles (which is possible through a fusion of protons and electrons into neutrons, via inverse beta-decay). When he insisted that the state-space of possibilities gave out, Lewis was assuming the constraint imposed by the gravitational attraction—under those conditions, and given PEP, for the star to occupy a state corresponding to a smaller radius is a physical impossibility for a star of a given number of fermions.

¹¹ In effect, the labels have to be permuted and an appropriate wave function then constructed. This permutation of the labels was, at the time, understood as signifying that the particles should not be regarded as individuals, although as it turns out, they can be albeit at a certain (metaphysical) cost; see French and Krause (2006).

The former explicitly applies PEP right at the start, to obtain what is known as the Slater determinant, in the case of fermions, where this describes the N-body wave function of the system, and from which one can then obtain a set of coupled equations for the relevant orbitals. The latter begins with the electron density in 3 spatial coordinates and via functionals of that density reduces the N-body problem of a system with 3N coordinates to one of 3 coordinates only. Again the technique explicitly incorporates the ‘exchange interaction’ due to PEP, and together valence bond theory and molecular orbital theory offer a complementary range of tools and techniques for describing and explaining various aspects of chemical bonding. Despite its name, exchange interaction (also sometimes called exchange force) is best construed as a purely *kinematical* consequence of quantum mechanics, having to do with the possible multi-particle wave functions allowed by PEP (or, more generally, Permutation Invariance).

For a specific illustration of the explanatory contribution of this kind of kinematic constraint, consider the solubility of salt. Examining the explanation of solubility brings out its non-causal character. We begin with the formation of an ionic bond between Na^+ and Cl^- , with the bond-dissociation energy (E_{diss}) measuring the strength of a chemical bond the breaking of which is required for the substance to dissolve:

$$E_{\text{diss}} = E^+ + E^- - \frac{Ke^2}{r} + C \frac{e^{-ar}}{r}$$

Here the first term stands for the ionization energy, the second for the electron affinity, the third for the Coulomb attraction, and the fourth describes the energy associated with the so-called ‘Pauli repulsion’, arising from PEP.¹² In this case, perhaps even more clearly than above, the sense of ‘repulsion’ is that of a *façon de parler*. The contribution of this symmetry-based term to the dissociation energy is critical, and it does not have a causal origin unlike the other terms, corresponding to none of the four known forces. Furthermore, there is no equivalent move available here to statistical abstraction, as in the case of quantum statistical ‘degeneracy pressure’.

Before we go on to analyse this explanation, it’s worth noting that examples of PEP-based explanations proliferate: numerous mechanical, electromagnetic, and optical properties of solids are explained by invoking PEP, including, indeed, the stability of matter itself.¹³ Perhaps in certain scenarios, such as that of the white dwarf collapse, a case can be made that the explanation involved can be accommodated within a broad causal (and, if this is the direction in which one’s metaphysical inclinations run, *dispositionalist*) framework. However, in the light of the wide range of explanations of very

¹² For the Pauli repulsion diagram for salt, see <<http://hyperphysics.phy-astr.gsu.edu/hbase/molecule/paulirep.html#c1>>.

¹³ For a quantum theoretic, PEP-based explanation of stability of matter, see e.g., Dyson and Lenard (1967, 1968). This was already anticipated by Fowler (1926), who only two years after Pauli’s proposal of his exclusion principle, suggested that PEP explains white dwarves’ stability.

different kinds of phenomena that turn on PEP (and the Permutation Invariance from which it is derived), we would argue that the recognition of the explanatory role played by this fundamental symmetry motivates a move beyond the causal schema to the framework of counterfactual dependence.

How, then, should we characterize these explanations? Let us begin by recalling that at the heart of the counterfactual-dependence view of explanation is the idea that an explanation proceeds on the back of some form of dependence between that which is described by the explanans and the phenomena captured by the explanandum. Strevens also considers, in this spirit, the example of the halting of white dwarf collapse and the role of PEP within his kairetic approach to explanation:

What relation holds between the law [PEP] and the arrest, then, in virtue of which the one explains the other? Let me give a partial answer: the relation is, like causal influence, some kind of metaphysical dependence relation. I no more have an account of this relation than I have an account of the influence relation, but I suggest that it is the sort of relation that we say “makes things happen”. (Strevens 2008: 178)

Metaphysically one can explicate this dependence in various ways (see French 2014), but what we regard as important with respect to the philosophy of explanation is that it can be cashed out via *counterfactual* dependence and thus can underwrite the appropriate counterfactual reasoning. Explanations, whether causal or non-causal, can be supported by a theory that correctly depicts a space of possible physical states with a sufficiently rich structure, such that it grounds robust reasoning that answers what-if-things-had-been-different questions.¹⁴ Such facts about state-space is precisely what we have in the white dwarf case, as Lewis noted. Similarly, in the explanation of salt’s solubility, and in a host of other explanations, PEP imposes a global constraint upon a space of possible physical states, yielding the robust explanatory dependence of the explanandum on the global symmetry. Due to the *global* character of that constraint the relevant counterfactuals are quite different from the interventionist counterfactuals associated with causal explanation. But the spirit of the Woodwardian counterfactual framework still holds.

In the case of PEP, the relevant counterfactuals involving changes in the explanans turn on asking ‘what if PEP did not apply?’ Note that what we have here is a ‘contra-nomic’ counterfactual (lumping laws and symmetries together for these purposes). There are, of course, a number of significant issues associated with how we evaluate such counterfactuals but which we do not have the space to go into here. Instead we shall limit ourselves to explicating it, and answering the question, in the context of our concrete examples.

In the case of the explanation of the solubility of salt, if PEP did not apply, then the crucial ionic bond would not form in the first place and we would not have any salt to

¹⁴ See Saatsi (forthcoming) for examples of explanations where the relevant structure of the space of possible states concerns closed loops (holonomies) in state space.

begin with! More fundamentally, if PEP did not apply then that would imply that electrons would not be fermions and we would not even have ions of sodium and chlorine because there would not be the constraint that leads to electrons occupying the relevant energy states in the way that underpins ionization (or, indeed, the formation of atoms!). In the case of the white dwarf, if PEP did not apply—namely, if the particles involved were not fermions—the Symmetrization Postulate dictates that the relevant quantum mechanical wave function must be symmetrized, yielding Bose-Einstein statistics. Of course, under that form of statistics the white dwarf collapse would not halt at all; indeed, what we would end up with is a form of ‘Bose-Einstein condensate’. For phenomena for which the requirement of symmetric wave functions is appropriate, the Symmetrization Postulate serves as an explanans for a whole host of different phenomena, from lasers to superconductivity and the ‘fountain effect’ in liquid helium-4, where very small temperature differences lead to dramatic (and ultimately non-classical) convection effects (see Bueno et al. 2002). And we can go further: if we replace the Symmetrization Postulate with the arguably even more fundamental requirement of Permutation Invariance, then, with the possibility of paraparticle statistics, we get a whole host of counterfactuals—indeed an infinite number—rather than just two. Here, quite interesting statistical behaviour emerges if we ask ‘what if there were para-particles of order such-and-such?’ for example. Or more generally perhaps, ‘what if we have deviations from either Bose-Einstein or Fermi-Dirac statistics?’ (see, for example, Greenberg 1992).¹⁵ And we can go further still: as already noted, in spaces of less than three dimensions, one can obtain kinds of particles (or, rather, ‘quasi-particles’) known as anyons,¹⁶ which explain the fractional quantum Hall effect, regarded as representing a new state of matter manifesting so-called ‘topological order’.¹⁷

To sum up, we have argued that in connection with explanations turning on fundamental discrete symmetries such as PEP we can avail ourselves of a counterfactual framework, but drop the requirement of interventions that effectively mark a causal dependence. What distinguishes the kinds of explanations we are concerned with from causal ones is the nature of the explanans. The relevant counterfactuals are theoretically well-formed (in the sense of being grounded in the relevant—mathematically described—physics), and if true they are indicative of dependence relations that hold between various explananda and fundamental symmetries of the world. But these dependence relations are not causal by virtue of involving a *global kinematic*

¹⁵ So, returning to the example of salt, we might ask, not just ‘what if electrons were bosons?’, in which case what we call ‘matter’ would look and behave very differently indeed (!), but ‘what if electrons were paraparticles of some order?’ In that case, not everything would degenerate into a Bose-Einstein condensate and quite interesting statistical behaviour would result. The point is, however, that changing the explanans would yield very different consequences.

¹⁶ As already noted in footnote 4, these are described by the ‘braid’ group which generalizes the permutation group.

¹⁷ Anyons are described as ‘quasi-particles’ since it remains contested whether they should be regarded as effectively mathematical devices or real; an experiment supposedly demonstrating the latter remains controversial (Camino et al. 2005). However, further suggestions have been made involving the experimental manipulation of anyons (see Keilmann et al. 2011).

constraint on the available physical states—an explanans for which the notion of intervention seems inapplicable.

We will bring our discussion of discrete symmetries to a close by suggesting that this analysis can also be extended to cases other than Permutation Invariance. One example is the explanation of universality of critical phenomena, which arguably crucially involves a non-causal dependence between specific universality classes, on the one hand, and a discrete symmetry property of the micro-level interactions (the symmetry of the ‘order parameter’), on the other. This dependence is brought out by renormalization group analyses of statistical systems (Reutlinger 2016). For another example, consider the so-called CPT Theorem and the explanations that invoke it. The theorem states that all Lorentz-invariant quantum field theories must also be invariant under the combination of charge conjugation (swapping + for – charges and vice versa; i.e., swapping matter for anti-matter), parity reversal (reflection through an arbitrary plane or flipping the signs of the relevant spatial coordinates of the system), and time reversal (flipping the temporal coordinate). It has been invoked to prove the Spin-Statistics Theorem, which states that particles that obey Bose-Einstein statistics must have integral spin and those that obey the Fermi-Dirac form must have half-integral spin.¹⁸ Violations of the components of the invariance also feature in scientific and philosophical explanations. For example, violation of CP symmetry has been used to explain the preponderance of matter in the universe, rather than an equal distribution of matter and anti-matter as would be expected. Our hunch is that such explanations also involve assumptions about non-causal counterfactual dependencies, but we shall not pursue this further here.

4. Continuous Symmetries

Let’s now move on to consider the other significant kind of symmetry found in science, continuous symmetries, and explanations they can support. Continuous symmetries are described by continuous groups of transformations (in particular the Lie groups which cover smooth differentiable manifolds and which underpin Klein’s ‘Erlangen’ programme of systematizing geometry). They are embodied in classical claims regarding the homogeneity and isotropy of space and the uniformity of time, and are accorded fundamental primacy over the relevant laws in the context of Special Relativity, where the Lorentz transformations are effectively promoted to universal, global continuous spacetime symmetries. The extension of such symmetries beyond the spacetime context, to the so-called local ‘internal’ symmetries in the context of fundamental interactions represents one of the major developments in physics of the past hundred years or so, underpinning the so-called Standard Model (see, for example, Martin 2003).

¹⁸ And likewise for parastatistics, since we’ve mentioned them.

One of the most celebrated explanatory uses of such continuous symmetries appeals to Noether's famous theorem, connecting continuous symmetries to the existence of conserved quantities. The issue of how to interpret that connection has been the subject of some debate. Thus, although many scientists and philosophers regularly speak of conservation laws being explained by symmetries or by Noether's theorem itself, some have challenged this idea. Brown and Holland (2004), for example, point to the two-way nature of Noether's (first) theorem: it not only allows for a derivation of conserved quantities from dynamical symmetries, but equally for the derivation of dynamical symmetries from knowledge of which quantities are conserved:¹⁹

[The] theorem allows us to infer, under ordinary circumstances for global symmetries, the existence of certain conserved charges, or at least a set of continuity equations. The symmetry theorem separately allows us to infer the existence of a dynamical symmetry group. We have now established a *correlation* between certain dynamical symmetries and certain conservation principles. Neither of these two kinds of thing is conceptually more fundamental than, or used to explain the existence of, the other (though as noted earlier if it is easier to establish the variational symmetry group, then a method for calculating conserved charges is provided). After all, the real physics is in the Euler–Lagrange equations of motion for the fields, from which the existence of dynamical symmetries and conservation principles, if any, jointly spring. (Brown and Holland 2004: 1138)

Lange (2007: 465) concurs that “it is incorrect to appeal to Noether's theorem to secure these explanations”, also pressing the point about the theorem's two-way directionality: “The link that Noether's theorem captures between symmetries and conservation laws is (ahem!) symmetric and so cannot account for the direction of explanatory priority.” Lange does not conclude that continuous symmetries cannot play an explanatory role, however, as he goes on to provide his own ‘meta-laws’ account of the modal hierarchy of symmetries and conservation laws with the intention to secure the explanatory priority of symmetries. We will comment on this account in due course, but let's first consider further the two-way directionality of Noether's theorem.

In our view—from the counterfactual-dependence perspective—little hangs on the fact that Noether's theorem represents a *correlation* between symmetries and conserved quantities. After all, most explanations in physics appeal to regularities that can underwrite derivations running in two directions, only one of which may be considered explanatory. (Our toy example in section 2 is a case in point, reflecting a point already familiar from explanations of flagpole shadows, pendulum periods, and so on.) What matters, rather, is whether the physics that connects symmetries and conserved quantities can be regarded as uncovering genuine (causal or non-causal) *dependencies* that underwrite explanations in which symmetries function as explanans. If this can be

¹⁹ Here we will only focus on Noether's first theorem, which relates conserved quantities to continuous (global) symmetries in Lagrangian dynamics. The second theorem has to do with local symmetries (namely, symmetries that depend on arbitrary functions of space and time; see Brading and Brown 2003).

done, then we can regard such dependencies as the source of the explanatory power of continuous symmetries.

This can be done. To show how, we will first recall the relevant theoretical context. (For details, see e.g., Neuenschwander 2011.) Noether's theorem concerns physical systems amenable to a description within Lagrangian dynamics, in which the system can be associated with a Lagrangian: a function of the system's configuration variables and their rate of change. The system's dynamical behaviour over time is such that it minimizes a functional of the Lagrangian over time. For a system in classical mechanics, for instance, this functional is the time integral of the difference between the kinetic and potential energies:

$$J = \int_a^b (K - U) dt = \int_a^b L dt$$

The requirement that the system's actual dynamics follows a trajectory that minimizes this functional is called Hamilton's principle. The coordinates of this trajectory will satisfy differential equations called Euler-Lagrange equations.

$$\frac{\partial L}{\partial x^\mu} = \frac{d}{dt} \frac{\partial L}{\partial \dot{x}^\mu}$$

In Lagrangian dynamics there are significant connections between symmetries and conserved quantities that flow out directly from the Lagrangian, without at all having to consider Noether's theorem (and the more general connection between conserved quantities and symmetries of the functional). For instance, it is a straightforward corollary of the Euler-Lagrange equations that canonical momentum p^μ is constant if and only if $\frac{\partial L}{\partial x^\mu} = 0$.²⁰ Similarly, it follows directly from the Euler-Lagrange equations that a system's Hamiltonian is constant if and only if the Lagrangian does not explicitly depend on t , viz. $\frac{\partial L}{\partial t} = 0$. When the Hamiltonian (formally defined as $H = p_\mu \dot{x}^\mu - L$) can be identified with (the numerical value of) the system's energy, it can thus be seen that energy conservation is connected to symmetry under a time translation.

These elementary connections between continuous symmetries and conserved quantities in the Lagrangian framework can be viewed as special cases of Noether's theorem, which in its full generality need not come into play in deriving the conserved quantities for a given Lagrangian.²¹ Based on these connections, mathematical derivations can run in reverse, too, so as to establish symmetries of the Lagrangian from a given set of conserved quantities. Again, these connections in and of themselves say nothing about explanatory priority. In order to get a handle on that we need to

²⁰ Canonical momentum is defined as $p_\mu = \frac{\partial L}{\partial \dot{x}^\mu}$ for each coordinate x^μ and its coordinate velocity $\dot{x}^\mu$.

²¹ Noether's theorem is broader in that it relates conserved quantities to the symmetries of the *functional* (not just the Lagrangian), yielding conserved quantities that are linear combinations of H and p_μ .

consider the modal information provided by the physics. From the perspective of the counterfactual-dependence account, this explanatory priority is underwritten by the fact that in a typical application of these results to a particular system (e.g., the solar system) there is a natural sense in which the conserved quantities *depend on* the features of the system represented by the Lagrangian and its symmetries, but not the other way around. The Lagrangian and its properties reflect the relevant properties of the system being described: kinetic and potential energy functions, and whatever constraints there are to its dynamics. When we consider changes to these features of the system, we consider changing, for example, the spatial distribution of mass or charge, or their quantity. These changes can have an effect on regularities manifested by the system as it evolves over time: different features of the system may become constants of motion, properties whose values are unchanged over time. The point is that there is no way to alter these regularities concerning the system's behaviour—these constants of motion—directly as it were, without acting upon the features of the system that determine the system's behaviour. And it is the latter that feature in the Lagrangian, the symmetries of which thereby determine the constants of motion in a way that supports explanatory what-if-things-had-been-different counterfactuals.

This asymmetry is best illustrated with a concrete example. For an elementary case, consider a particle moving under a central force. In spherical coordinates (r, θ, φ) , the potential energy $U(r)$ of the particle depends only on the radial coordinate r , when a spherically symmetric source of e.g., gravitational or electric force field is located at the origin. The kinetic energy function

$$K = \frac{1}{2}m\dot{\mathbf{v}}^2 = \frac{1}{2}m(\dot{r}^2 + r^2\dot{\theta}^2 + r\dot{\varphi}^2 \sin^2 \theta)$$

feeds into the Lagrangian $L = K - U(r)$. From Euler-Lagrange equations we get as (separate) constants of motion the azimuthal and polar components of the orbital angular momentum: $p_\theta = mr^2\dot{\theta}$ and $p_\varphi = mr^2\dot{\varphi} \sin^2 \theta$. This is why the particle's trajectory is constrained to a plane; this regularity about the dynamics depends on the symmetry of the Lagrangian (namely, symmetry of kinetic and potential energy functions).

Changing the potential energy function, either in its strength (by varying the amount of mass or charge at the centre), or in its spatial geometry by breaking the spherical symmetry in favour of some other symmetry, will have effects on the dynamical behaviour of bodies moving under the potential. These effects are reflected also in the regularities of the dynamics captured by the constants of motion. Grasping the connection between these constants of motion and the symmetries of the Lagrangian enables us to answer what-if-things-had-been-different questions such as: What if the source were not spherically symmetrical? What if the source were a spheroid, as opposed to a sphere? What if the spheroid revolved about its minor axis? What if it oscillated in a particular way? From the counterfactual-dependence perspective

this kind of modal information is explanatory: it places the explanandum in a pattern of counterfactual dependencies (as Woodward puts it), thus bringing out how the regular aspects of the dynamics captured by the conserved quantities depend on the symmetries.²²

In this simple example the asymmetry of dependence is amenable to a ‘manipulationist’ interpretation, given that the notion of intervention is applicable to the relevant features of the central force system that function as the explanans (cf. Woodward 2003). However, it is worth noting that the explanandum is a *regularity*, and it’s not clear whether there is a corresponding event explanandum at all. This casts some doubt on whether the explanation in question should really count as causal (Saatsi and Pexton 2013). Furthermore, we will now argue that such an interventionist interpretation of symmetry *qua* explanans need not always be available, and even if it is not available, the derivation of conserved quantities from symmetries can nevertheless be explanatory.

In particular, assume that the closed system we are concerned with is the *whole* universe with its dynamical laws, represented via the Lagrangian, exhibiting certain symmetries. We can, again, answer counterfactual questions of the sort ‘*What if the universe were not symmetrical in this or that way?*’ Answers to such what-if-things-had-been-different questions bring out the way in which particular conservation laws are counterfactually related to the symmetries at stake, even though it is not clear that counterfactuals regarding alternative symmetries can be interpreted in causal terms, with reference to possible manipulations or interventions. The global symmetries of dynamical laws seem intuitively on a par with e.g., the dimensionality of space—a global feature which Woodward once mooted as grounding a non-causal counterfactual-dependence explanation of the stability of planetary orbits (Woodward 2003: §5.9).

One might worry that we do not have a sufficiently solid grasp on the sense of counterfactual ‘dependence’ between the symmetries of dynamical laws and conservation laws. Why dependence, as opposed to a mere correlation, as Brown and Holland suggest? We think the reason that physicists often give explanatory priority to symmetries over conservation laws has to do with the fact that in analogous applications of Noether’s theorem to particular *subsystems* of the universe, such as the central-force system examined above, the explanatory priority is transparent, partly due to the applicability of notions of manipulation and interventions. Explanatory reasoning about the relationship between conserved quantities and symmetries is naturally extended from such subsystems, involving e.g., central or harmonic forces, to symmetries of the laws covering the whole universe. Given the tight connection between conserved quantities and continuous symmetries in the Lagrangian framework—a connection which Noether’s theorem captures in highly general terms—we

²² This is analogous to the connection between a gravitational pendulum’s length and its period. For a given pendulum, we can explain a feature of its dynamical behaviour over time, namely its period, in terms of its length (and the gravitational potential). But we do not explain the pendulum length in terms of the period, even though the pendulum law allows for its derivation.

naturally understand and explain conservation laws in terms of symmetries. This provides a non-causal explanation of particular conservation laws, capturing pervasive regularities of dynamical systems.

The explanatory dependence appealed to here need not be a matter of deep metaphysics. Indeed, insofar as our understanding of the counterfactual-dependence analysis of explanation is concerned, this perspective is meant to be compatible with both Humean and non-Humean approaches to the metaphysics of modality and laws. Remember that for the Humean, dynamic and conservation laws alike are just statements of worldly regularities, the special status of which is underwritten by features of the whole global ‘mosaic’ of particular facts. Understanding the law-like status of those regularities is partly a matter of grasping which features of the mosaic are responsible for that special status. For the regularities involving conserved quantities, the relevant features involve symmetries, statements regarding which would feature as axioms in the relevant formalization, according to the Best System Analysis of laws. Grasping how those symmetries are responsible for the regularities that conservation laws represent is only a matter of seeing how mosaics with different symmetries would yield different conservation laws. For the Humean there is no deeper metaphysical connection between symmetries and conservation laws: both concern regularities of the mosaic, connected by Noether’s theorem. The connection is necessary to a stronger degree of necessity than nomological or causal necessity, and as such comparable to ‘distinctly mathematical’ explanations (Lange 2013; Jansson and Saatsi forthcoming).²³ Admittedly there is much more to be said to elaborate on this sketch, and the nature of conservation laws and symmetries is a largely unexplored area of Humean metaphysics of science.²⁴

Alternatively, one could try to accommodate such symmetries within a dispositionalist approach to modalities and laws. Bird (2007) dismisses symmetries as temporary features of science, to be dropped from our metaphysics as science progresses. And certainly, the prospects for capturing symmetries via the standard stimulus-and-manifestation characterization of dispositions look dim (see French forthcoming). Nevertheless, one might adapt some of the recently proposed metaphysical devices in this area to articulate an account of how symmetries might be understood as obtaining from a powers-based metaphysics (see Vetter 2015). More plausibly, perhaps, if one were to insist on giving modality some metaphysical punch, as it were, one could interpret symmetry principles such as Permutation Invariance as ‘encoding’, in a sense, the relevant possibilities. By virtue of that, they could then be understood as inherently or, perhaps, primitively, modal. If, further, such principles were taken to be features of the structure of the world, one would reach a position that could be considered a ‘third way’ between Humean and dispositionalist accounts (French 2014). And of course, on such a view, the role of such

²³ See also Saatsi and Reutlinger (forthcoming) for a related point of view on renormalization group explanations.

²⁴ For a significant exception, see Yudell (2013).

principles as the explanans in the kinds of explanations we have considered here would correspond with their ontological priority as such structural features.

However, our central point about the explanatory character of symmetry explanations is meant to be independent of the metaphysics of modality that underwrites the explanatory counterfactuals that, in turn, answer the relevant what-if-things-had-been-different questions. This is in contrast with Lange (2007), who regards symmetry principles in science as deeper *meta*-laws that constrain the laws there could be: given such meta-laws, the range of possible laws is restricted to those that comply with the symmetry principles in question. Lange motivates this anti-Humean metaphysics of (meta-)laws by drawing on utterances from prominent scientists, such as Feynman:

When learning about the laws of physics you find that there are a large number of complicated and detailed laws, laws of gravitation, of electricity and magnetism, nuclear interactions, and so on, but across the variety of these detailed laws there sweep great general principles which all the laws seem to follow. Examples of these are the principles of conservation. All the various physical laws obey the same conservation principles. (Feynman 1967: 59)

Although we are sympathetic with the naturalistic spirit of Lange's programme, we also see it as potentially question-begging against the competing Humean accounts. From the perspective of a non-governing conception of laws, the idea that laws are *governed* by higher symmetry principles is obviously problematic, to say the least. Furthermore, arguably the Humean has an alternative account to offer, as indicated above. As far as symmetries are employed to explain *particular* law-like regularities, including specific conservation laws, we maintain that this can be captured in the counterfactual-dependence framework.

Admittedly one can ask a deeper question of why various laws are unified in such a way that they are seemingly *governed* by one and the same symmetry principle. (For example, why are Newton's gravitational law and *Coulomb's* law *both* symmetric under arbitrary spatial displacement?) But although answers to this question are probably not amenable to counterfactual-dependence treatment, it seems to us that the question may not have a scientific explanation at all. Lange provides one *meta-physical* answer to it, Humeans offer another, and structural realists yet another. Assessment of the respective vices and virtues of these competing answers is a matter of wholesale comparison of 'metaphysical packages', and must be left for another occasion. Let us just say that appealing to scientists' sense of 'governance' at the level of broad symmetry principles and meta-laws is potentially question-begging in the way such appeal has been deemed problematic at the level of laws 'governing' events and regularities (Beebe 2000).

We will not pursue this metaphysical issue further here, but instead comment on Lange's take on Noether's theorem. According to Lange, Noether's theorem is irrelevant for explaining conservation laws. The argument partly turns on the noted symmetry of the theorem, already discussed above, and partly on the fact that "explanations [of conservation laws] were given long before anything resembling Noether's theorem

had been even remotely stated" (2007: 465). Lange is right to note this, of course, and we also emphasized the fact that in Lagrangian dynamics symmetries can be linked to conserved quantities in straightforward ways that do not demand anything like the full generality of Noether's theorem. Having said this, it seems to us that Noether's theorem is nevertheless explanatorily relevant in the following sense: it functions in a way analogous to an extremely broad-ranging invariant generalization in supporting counterfactual reasoning, by providing a link between symmetries and conservation that enables us to answer what-if-things-had-been-different questions for a maximal range of alternative situations. As such, the explanatory relevance of Noether's theorem is comparable to that of Euler's mathematical proof (regarding the necessary and sufficient conditions for a graph to have an Eulerian circuit) in relation to the impossibility of traversing all Königsberg's bridges by crossing each only once. In both cases we could in principle appeal to much more narrow-ranging generalizations connecting the relevant variables, but the respective mathematical theorems have maximal generality. (Cf. Jansson and Saatsi forthcoming for related discussion of the Königsberg's case.)

5. Conclusion

We started our discussion of symmetry explanations with an exceedingly simply toy example, a balance remaining in a state of equilibrium, which was explained by a symmetry of the forces involved. The more interesting real-life symmetry explanations discussed thereafter vary in their features, involving: discrete vs. continuous symmetries; local vs. global symmetries; symmetries that are fundamental vs. non-fundamental. Despite this variance, the cases we have discussed are unified in their explanatory character, which, we have argued, is naturally captured in the counterfactual-dependence framework.

Acknowledgements

Thanks to Callum Duguid for discussions of Humean approaches to symmetries, and to Alex Reutlinger and Jim Woodward for helpful comments.

References

- Beebe, H. (2000), 'The Non-Governing Conception of Laws of Nature', *Philosophy and Phenomenological Research* 61: 571–94.
- Bird, A. (2007), *Nature's Metaphysics: Laws and Properties* (Oxford: Oxford University Press).
- Brading, K. and Brown, H. R. (2003), 'Symmetries and Noether's Theorems', in K. Brading and E. Castellani (eds.), *Symmetries in Physics: Philosophical Reflections* (Cambridge: Cambridge University Press), 89–109.
- Brading, K. and Castellani, E. (eds.) (2003), *Symmetries in Physics: Philosophical Reflections* (Cambridge: Cambridge University Press).

- Brown, H. R. and Holland, P. (2004), 'Dynamical versus Variational Symmetries: Understanding Noether's First Theorem', *Molecular Physics* 102: 1133–9.
- Bueno, O., French, S., and Ladyman, J. (2002), 'On Representing the Relationship between the Mathematical and the Empirical', *Philosophy of Science* 69: 497–518.
- Camino, F. E., Zhou, W., and Goldman, V. J. (2005), 'Realization of a Laughlin Quasiparticle Interferometer: Observation of Fractional Statistics', *Physics Review B* 72: 075342.
- Carson, C. (1996), 'The Peculiar Notion of Exchange Forces I: Origins in Quantum Mechanics, 1926–1928', *Studies in History and Philosophy of Modern Physics* 27: 23–45.
- Dyson, F. J. and Lenard, A. (1967), 'Stability of Matter, Part I', *Journal of Mathematical Physics* 8: 423–34.
- Dyson, F. J. and Lenard, A. (1968), 'Stability of Matter, Part II', *Journal of Mathematical Physics* 9: 698–711.
- Feynman, R. P. (1967), *The Character of Physical Law* (London: Penguin).
- Fowler, R. H. (1926), 'On Dense Matter', *Monthly Notices of the Royal Astronomical Society* 87: 114–22.
- French, S. (1995), 'The Esperable Uberty of Quantum Chromodynamics', *Studies in History and Philosophy of Modern Physics* 26: 87–105.
- French, S. (2014), *The Structure of the World* (Oxford: Oxford University Press).
- French, S. (forthcoming), 'Doing Away with Dispositions', in A. Spann and D. Wehinger (eds.), *Dispositionalism: Perspectives from Metaphysics and the Philosophy of Science* (Dordrecht: Springer).
- French, S. and Krause, D. (2006), *Identity in Physics: A Historical, Philosophical, and Formal Analysis* (Oxford: Oxford University Press).
- French, S. and Rickles, D. (2003), 'Understanding Permutation Symmetry', in K. Brading and E. Castellani (eds.), *Symmetries in Physics: Philosophical Reflections* (Cambridge: Cambridge University Press), 212–38.
- Friedman, M. (1974), 'Explanation and Scientific Understanding', *Journal of Philosophy* 71: 5–19.
- Gavroglu, K. (1995), *Fritz London: A Scientific Biography* (Cambridge: Cambridge University Press).
- Greenberg, O. W. (1990), 'Example of Infinite Statistics', *Physical Review Letters* 64: 705.
- Greenberg, O. W. (1992), 'Interactions of Particles Having Small Violations of Statistics', *Physica A: Statistical Mechanics and its Applications* 180: 419–27.
- Jansson, L. and Saatsi, J. (forthcoming), 'Explanatory Abstractions', *British Journal for the Philosophy of Science*. <https://doi.org/10.1093/bjps/axx016>.
- Keilmann, T., Lanzmich, S., McCulloch, I., and Roncaglia, M. (2011), 'Statistically Induced Phase Transitions and Anyons in 1D Optical Lattices', *Nature Communications* 2, Article number 361. DOI: 10.1038/ncomms1353.
- Kitcher, P. (1981), 'Explanatory Unification', *Philosophy of Science* 48: 507–31.
- Kitcher, P. (1989), 'Explanatory Unification and the Causal Structure of the World', in P. Kitcher and W. Salmon (eds.), *Scientific Explanation* (Minneapolis: University of Minnesota Press), 410–505.
- Lange, M. (2007), 'Laws and Meta-Laws of Nature: Conservation Laws and Symmetries', *Studies in History and Philosophy of Modern Physics* 38: 457–81.
- Lange, M. (2012), 'There Sweep Great General Principles Which All the Laws Seem to Follow', *Oxford Studies in Metaphysics* 7: 154–85.

- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctively Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Lewis, D. K. (1986), 'Causal Explanation', in *Philosophical Papers*, vol. II (New York: Oxford University Press), 214–40.
- Martin, C. (2003), 'On the Continuous Symmetries and the Foundations of Modern Physics', in K. Brading and E. Castellani (eds.), *Symmetries in Physics: Philosophical Reflections* (Cambridge: Cambridge University Press), 29–60.
- Messiah, A. M. L. and Greenberg, O. W. (1964), 'Symmetrization Postulate and Its Experimental Foundation', *Physical Review* 136: B248–67.
- Neuenschwander, D. E. (2011), *Emmy Noether's Wonderful Theorem* (Baltimore, MD: Johns Hopkins University Press).
- Olver, P. J. (1995), *Equivalence, Invariants, and Symmetry* (Cambridge and New York: Cambridge University Press).
- Pincock, C. (2007), 'A Role for Mathematics in the Physical Sciences', *Noûs* 41: 253–75.
- Pincock, C. (2014), 'Abstract Explanations in Science', *British Journal for the Philosophy of Science* 66: 857–82.
- Redhead, M. L. G. (1984), 'Unification in Science', *British Journal for the Philosophy of Science* 35: 274–79.
- Reutlinger, A. (2016), 'Is There a Monist Theory of Causal and Non-Causal Explanations? The Counterfactual Theory of Scientific Explanation', *Philosophy of Science* 83: 733–45.
- Saatsi, J. (forthcoming), 'On Explanations from "Geometry of Motion"', *British Journal for the Philosophy of Science*. DOI: 10.1093/bjps/axw007.
- Saatsi, J. and Pexton, M. (2013), 'Reassessing Woodward's Account of Explanation: Regularities, Counterfactuals, and Noncausal Explanations', *Philosophy of Science* 80: 613–24.
- Saatsi, J. and Reutlinger, A. (forthcoming), 'Taking Reductionism to the Limit: How to Rebut the Antireductionist Argument from Infinite Limits', *Philosophy of Science*.
- Skow, B. (2014), 'Are There Non-Causal Explanations (of Particular Events)?', *British Journal for the Philosophy of Science* 65: 445–67.
- Strevens, M. (2008), *Depth: An Account of Scientific Explanation* (Cambridge, MA: Harvard University Press).
- van Fraassen, Bas (1991), *Quantum Mechanics: An Empiricist View* (Oxford: Oxford University Press).
- Vetter, B. (2015), *Potentiality: From Dispositions to Modality* (Oxford: Oxford University Press).
- Woodward, J. (2003), *Making Things Happen: A Causal Theory of Explanation* (Oxford: Oxford University Press).
- Yudell, Z. (2013), 'Lange's Challenge: Accounting for Meta-Laws', *British Journal for the Philosophy of Science* 64: 347–69.

10

The Non-Causal Character of Renormalization Group Explanations

Margaret Morrison

1. Introduction

One of the most commonly cited instances of non-causal explanation is mathematical explanation. The defining characteristic of the latter is that explanatory information comes via mathematics alone rather than from some combination of mathematical and other qualitative facts. The problem of determining exactly how mathematics can function in this way has been extensively discussed in the literature (Baker 2005, 2009; Bangu 2012; Batterman 2010; Lange 2013; Pincock 2007, 2012; Steiner 1978, to name a few). Rather than address specific features of these various arguments I want to draw attention to the type of mathematical explanation provided by renormalization group (RG) methods. Batterman's work has been influential in addressing the role of RG techniques and highlighting the type of non-causal information they provide. More recently, Reutlinger (2014) has also discussed these issues. My treatment here represents a somewhat different approach than Batterman's and Reutlinger's in that it stresses how the application of RG methods to dynamical systems more generally, as well as the relation between RG and probability theory, illustrates exactly how these explanations are non-causal.

Part of my argument is that the non-causal character of RG explanations is not due simply to the elimination of microscopic information resulting from the iterative application of the transformation. Instead, it is the role of fixed points together with the specific way RG acts on the structural features of the system (as represented in the Hamiltonians) that provide a physical, non-causal understanding of its behaviour. An important consequence of the evolution produced by RG transformations is not just that appeals to micro-foundations as sources of causal information are eliminated but rather that the explanation of universal behaviour *cannot* be given in terms of the system's interacting parts. What the RG framework does is transform a problem from one that incorporates specific model-based solutions to one based on generalized rules for

treating different kinds of dynamical systems, not just universality classes associated with phase transitions in statistical physics.¹

One might want to object here that renormalization group techniques are simply calculational tools and that explanations in statistical physics involving phase transitions and universality classes are typically going to appeal to probabilistic features that often embody causal information.² However, as we shall see later in this chapter, RG explanations aren't probabilistic in the usual sense. And, the way they differ from ordinary statistical mechanical explanations exemplifies why they are strongly non-causal.

I begin with a brief discussion of some of the contemporary views on non-causal, mathematical explanation, as well as some preliminary claims about why RG should be considered an instance of this. In section 3 I briefly discuss the issues related to phase transitions and the problems associated with micro-causality and probabilistic averaging, features that typically figure in explanations in statistical mechanics. From there I go on to address specific aspects of the non-causal, structural character of RG explanations and the relationship between RG and probability theory. Again, this feature of the argument is crucial for the claim that the non-causal status of RG explanations involves more than simply ignoring or "averaging over" microphysical details. I conclude with a discussion of the role of RG in dynamical systems and how that role exemplifies not only the structural aspects of RG explanations but how that structure also exemplifies the non-causal features. Each of the steps in the argument puts forward reasons why RG explanations should be considered non-causal. While each claim is to some degree autonomous, together they present what I see as a comprehensive picture of exactly how RG provides non-causal, but nevertheless physical, information.

2. Non-Causal, Mathematical Explanation—Some Background

What, exactly, does it mean to explain an event or a thing without citing the causes for its occurrence? Although there are many competing accounts of causality, and indeed causal explanation (mechanistic, manipulation, probabilistic, Humean regularity, etc.), various difficulties with these approaches have not dampened the intuitive appeal causal explanation seems to enjoy. We often feel we don't fully understand things until we know their causes, something that seems true regardless of whether our favoured "theory" of causality can be successfully defended. Sometimes we can cite X as the cause of Y even though we aren't exactly sure of the details of the chain of events responsible for Y being brought about by X, as in the case of diseases. Generally, we

¹ I will have more to say about how these "generalized rules" function in the discussion below.

² Universality classes are classes of phenomena that have radically different microstructures, like liquids and magnets, but exhibit the same behaviour at critical point.

think of an explanation as causal if it cites a condition(s), mechanism, or entity that is responsible for producing a new state of affairs or changing an existing one in some specific way. The cited cause may, but need not, be necessary; it could simply be a contingent feature of the state of affairs in question. Despite its rather widespread appeal, causal explanation is by no means a straightforward issue. Whether or not an explanation is causal can depend not only on what is being explained but also what counts, in the context, as a cause. For example, should we count background conditions as properly causal insofar as they constitute necessary enabling conditions? My intention is not to engage these debates here; one needn't have a well worked out theory of causal explanation in order to identify when an explanation conveys information *without* appeal to causes. However, we need to exercise caution since what might appear to be a non-causal explanation sometimes relies on causal factors. The example below is just such a case.

If we consider typical explanations in physics (insofar as that is possible), they usually proceed by invoking laws or law-like generalizations to infer future states from past ones. The state of the system at T_1 can be specified by the dynamical state of its constituents and this state generates, via laws of dynamics, a future dynamical state of the system that is characterized by the micro-constituents. It is relatively unproblematic to refer to this as an instance of causal explanation despite the fact that it is often a statistical relation that is being described. But now consider another example: the ideal gas law which states that $PV = RT$. Initially this appears to simply specify functional relations among quantities at a particular time, saying nothing about a time evolution or other causal features. If we ask why a particular gas has a specific volume we can answer the question by specifying the law and the values of the other quantities. However, the reason these relationships hold depends on the molecular structure of the gas; the law presupposes we have molecules that are infinitesimal in size with no forces acting between them. Because these assumptions are required for the law to hold, we can, or indeed should, think of them as providing necessary causal background conditions. Should these conditions change, the law is no longer valid and a different one needs to be invoked. In that sense what looks like a non-causal explanation turns out to involve causal information.

One example in statistical mechanics (SM) that appears decidedly non-causal is the case of equilibrium explanations where the most probable micro condition is equilibrium. For systems with vast numbers of particles the most probable value (overwhelmingly) for an appropriate function of the micro conditions will be the mean value over all the possible micro conditions for the system. Sklar (1993) calls this a "transcendental deduction"—we assume a fundamental fact (the existence of equilibrium states) and then ask how such states are possible. While this is not obviously causal we still might want to argue that it is the micro conditions plus the probabilistic assumptions that are the basis for the explanation. The question then transforms into one of how to interpret probabilities in SM and whether they involve some kind of "causal" information.

I will come back to the issue of probabilities later but for now let me redirect the discussion to mathematical explanation, which many argue is a paradigm case of non-causal explanation. Again, there are competing accounts of what makes an explanation mathematical. Baker (2009) claims that all we need for a mathematical explanation is that the physical fact in question is explained by a mathematical fact or theorem. The now famous cicada example is an illustration. Two North American subspecies of cicadas spend 13 and 17 years underground in larval form. Why have the life cycles evolved into periods that correspond to prime numbers? Because having a life cycle period that minimizes intersection with other periods is evolutionarily advantageous and it is a theorem of number theory that prime periods minimize intersection.

Baker takes this to be an example of an *indispensable, mathematical explanation* of a *purely physical phenomenon*; in other words, the ‘mathematical’ features of the explanation are a necessary feature. Moreover, the indispensability of the mathematical features turns out not to be limited to cicadas; there are other explanations that rely on certain number theoretic results to show that prime cycles minimize overlap with other periodical organisms. Avoiding overlap is beneficial whether the other organisms are predators, or whether they are different subspecies since mating between subspecies would produce offspring that would not be coordinated with either subspecies.

But surely this explanation also has a causal element that is described by the biological information about these life cycles. One might want to argue that the underlying problem here is, of course, trying to separate what’s truly mathematical in the explanation from what’s physical or biological. And, indeed, it would seem that in this case the basis for the explanation is a law that combines mathematical and biological information. While the mathematics may be an indispensable part of the explanation it is not the sole explanatory factor. The evolutionary advantage in avoiding intersection with other periods provides us with a form of causal information that is also crucial in understanding the life cycle period. Hence, the indispensability of the mathematics here doesn’t seem to entail that the explanation is non-causal.

This interplay of physical (or biological, etc.) and mathematical information is a common problem in the attempt to give an account of how to characterize mathematical explanation, and whether those explanations can be properly classified as non-causal. Lange (2013) has an extremely persuasive discussion of these issues which culminates in his own account of when an explanation is truly mathematical and what the relation to causal (or non-causal) explanation is in these cases. Lange argues (2013: 487) that mathematical explanations are non-causal because they show how the fact to be explained was inevitable to a “stronger degree than could result from the causal powers bestowed by the possession of various properties”. In other words, the modal strength of the connection between causes and effects is insufficient to account for the inevitability of the explanandum. What Lange quite rightly points out is that an explanation is not deemed non-causal simply because it doesn’t appeal to causally active entities. Indeed, non-causal explanations can contain detailed causal histories and laws that do not function as explanatory factors. By contrast,

mathematical explanations work by constraining what's *possible* in much the same way that symmetry principles operate (2013: 495); in other words, they provide a set of structural constraints that govern the system in question.³ The important point is that despite allowing causal information into the explanans, any connection between the cause(s) and the explanandum holds not in virtue of a physical law but by *mathematical* necessity (2013: 497). In that sense the mathematics is what functions as the explanatory vehicle.

One example of this kind of explanation discussed by Lange concerns why there are at least four equilibrium configurations for a simple double pendulum. We can explain this causally by identifying the particular forces on the two bobs and then determining the configurations under which both have zero net force. By Newton's second law they will then undergo no acceleration and will remain at rest once they are in that configuration. Lange claims that this is a causal explanation but he also identifies a non-causal one for the same phenomenon, one that ignores the particular forces acting on the system and instead appeals only to the fact that in virtue of being a double pendulum the system's configuration space is the surface of a torus. And this is applicable to all double pendula, not just simple ones. Although it also appeals to a particular case of Newton's second law—that a system is in equilibrium when the net force on each of its parts is zero—Lange claims that this is a general constraint that applies regardless of what forces are acting. As he puts it: “mathematical explanation works [...] by showing how the explanandum arises from the framework that any possible causal structure must inhabit” (2013: 505).

An important feature of Lange's account is his claim that there is no criterion that sharply distinguishes a *distinctively* mathematical explanation from other non-causal explanations that appeal to mathematical facts (2013: 507). Instead, it is a matter of context and degree; and, where the mathematical facts alone are doing the explaining then the explanation is *distinctively* mathematical. Although Lange has provided a compelling argument for differentiating causal and mathematical (and non-causal) explanation the problem remains of determining (even in a given context) when, exactly, it is *just* the mathematics that is functioning as the explanatory vehicle. Because mathematical explanations will undoubtedly come in different forms it seems entirely reasonable, as Lange suggests, that we shouldn't expect to arrive at necessary and sufficient conditions for identifying them across the board.

That said, one might want to argue that the kind of general constraints specified by Lange as indicative of mathematical explanation can sometimes function in a causal manner. Certain types of symmetries are a case in point. The global gauge invariance of a phase determines charge as a conserved quantity while local invariance can be seen as determining the existence of the electromagnetic field under the form of Maxwell's equations. The interactions described by Maxwell's equations can then be

³ The latter interpretation in terms of structural constraints is mine, not Lange's, but I think it captures the spirit of his view.

said to furnish the causes of the observed effects. Hence, one can understand this in a hierarchical manner, with very general causal constraints given by the symmetries; constraints that provide generic causal explanatory information.

My discussion of RG as a type of mathematical (and non-causal) explanation is similar in spirit to Lange's in that it emphasizes very general features of systems. However, it differs in that the explanatory power comes not from the modal character of a law stated in mathematical terms but from the fact that RG is a particular type of mathematical framework used to explain structurally stable behaviour in physical systems. If we ask why certain types of systems undergoing phase transitions can be grouped into universality classes we pose a why-question but I claim that the answer does not involve importing causal information, even in the generic sense described above. In the case of RG there is no appeal to the underlying "physics" as the source of causal information. Although the symmetry and dimensionality of the system are important in these contexts, the symmetry considerations operate differently than the local or global symmetries associated with gauge or phase invariance mentioned above.

Reutlinger (2014) has also argued for the non-causal, mathematical aspects of RG explanations. He claims that neither of the two mathematical operations involved in RG explanations—the RG transformations on the Hamiltonians that enable physicists to ignore aspects of the interactions between micro components, and a "flow" or mapping of transformed Hamiltonians to the same fixed point—is best understood as directly revealing information about cause–effect relations. Reutlinger's point here is to challenge Batterman's (2010) claim that if an explanation ignores causal (micro) details, which RG explanations certainly do, then the explanation is non-causal. Instead, he claims that RG explanations are mathematical in virtue of the application of mathematical operations, which do not serve the purpose of representing causal relations.

Initially this sounds very similar to my own view but my argument differs in scope in that it emphasizes how the more general structural aspects of RG explanations serve to distinguish them from probabilistic approaches to explanation, and how, in virtue of this, they provide non-causal, physical information across a variety of contexts. This is important because it is crucial to distinguish between explanations that employ mathematical operations like statistical averaging, which also ignores specific causal details, and the kind of mathematical approach embedded in RG. The specific details of my differences with Reutlinger's account will become apparent in the discussion below, but now let me move on to a brief review of the RG methodology and its relation to the microphysics of statistical mechanics.

3. RG, Statistical Mechanics, and Micro-Causality

The theory of critical phenomena deals with continuous or second-order phase transitions in macroscopic systems (e.g., magnetic transitions and superfluid He) that occur under many different conditions such as wide temperature ranges (ferromagnetic at

1000K and Bose-Einstein condensation at 10^{-7} K). The theoretical account of phase transitions requires a mathematical technique known as taking the “thermodynamic limit” $N \rightarrow \infty$, sometimes called the infinite volume limit, where the volume is taken to grow in proportion with the number of particles while holding the particle density fixed. But why should we need to assume an infinite volume limit to explain, understand, and make predictions about the behaviour of a real, finite system? A defining characteristic of a phase transition is an abrupt change, noted mathematically by a singularity. In other words, thermodynamic observables characterizing the system are not defined or “well-behaved”—they are not differentiable. All thermodynamic observables are partial derivatives of the partition function, hence, a singularity in the partition function is required to obtain a singularity in the thermodynamic function. Although the partition function is analytic, it is possible for systems with infinite N to display singular behaviour for non-vanishing partition functions. The problem is that in SM the solutions to the equations of motion that govern these systems are analytic and as a result, there are no singularities and hence no basis for explaining phase transitions. Note that the problem here is not that the limit provides an easier route to the calculational features associated with understanding phase transitions; rather, the assumption that the system is infinite is *necessary* (theoretically) for the symmetry breaking associated with phase transitions to occur. In other words, we have a description of a physically unrealizable situation (an infinite system) that is *required* to explain a physically realizable phenomenon (the occurrence of phase transitions in finite systems).

One of the interesting features of phase transitions is that the effects of this singularity are exhibited over the entire spatial extent of the system; hence, the occurrence of a phase transition in these systems (infinite particles, volume or sometimes strong interactions) involves a variation over a vast range of length scales. From a mathematical perspective we can think of the RG as a technique that allows one to investigate the changes to a physical system viewed at different distance scales. To see how the process takes place let us look briefly at the real space RG approach that stems from the Wilson–Kadanoff method which involves scaling relations on a lattice of interacting spins (e.g., ferromagnetic transition) and transformations from a site lattice with the Hamiltonian $H_a(S)$ to a block lattice with Hamiltonian $H_{2a}(S)$.⁴

If one starts from a lattice model of lattice size a one can sum over degrees of freedom at size a while maintaining their average on the sub-lattice of size $2a$ fixed. Starting from a Hamiltonian $H_a(S)$ on the initial lattice one can generate an effective Hamiltonian $H_{2a}(S)$ on the lattice of double spacing. This transformation is repeated as

⁴ The contrast is with the momentum space approach initially put forward by Gell-Mann and Low (1954) for quantum field theory (QFT). Wilson (1971) transformed Kadanoff’s (1966) block spin method into a more precise computational scheme which eventually bridged the gap with the RG of QFT. He essentially used the momentum space description of the block spin picture to analyse the Ginzburg–Landau model, and extending the momentum space concept he solved the Kondo problem which dealt with the effect of magnetic impurity on the conduction band electrons in a metal. It was the first instance of a full implementation of the RG method. Several variants of the Wilson RG were later introduced in both momentum and real space.

long as the lattice spacing remains small compared to the correlation length. The key idea is that the transition from $H_a(S)$ to $H_{2a}(S)$ can be regarded as a rule for obtaining the parameters of $H_{2a}(S)$ from those of $H_a(S)$. The process is then repeated with the lattice of small blocks being treated as a site lattice for a lattice of larger blocks, with each block considered as a new basic entity. One then calculates the effective interactions between them and constructs a family of corresponding Hamiltonians. The coarse-graining process provides the bridge from the micro to the macro levels and each state in between. Moving from small to larger block lattices gradually excludes the small scale degrees of freedom such that for each new block lattice one constructs effective interactions and finds their connection with the interactions of the previous lattice. The iterative procedure associated with RG results in the system's Hamiltonian becoming more and more insensitive to what happens on smaller length scales, or as we saw above, the system losing memory of its microstructure. What this means is that the microphysics has been "transformed" via RG in a way that detaches it from the stable macro behaviour.

To see in a little more detail just how this works we need to show how the critical behaviour characteristic of a phase transition is expressed mathematically. The iterative application of the RG transformation is related to a scale invariance symmetry which enables us to see how and why the system appears the same at all scales (self-similarity). The symmetry of the phase transition is reflected in the order parameter (e.g., a vector representing rotational symmetry in the magnetic case, and a complex number representing the Cooper pair wave function in superconductivity), with a non-zero value for the order parameter typically associated with this symmetry breaking.

The correlation function $G(r)$ measures how the value of the order parameter at one point is correlated to its value at some other point. Usually, near the critical point ($T \rightarrow T_c$), the correlation function can be written in the form

$$G(r) \approx \frac{1}{r^{d-2+\eta}} e^{\left(\frac{-r}{\xi}\right)},$$

where r is the distance between spins, d is the dimension of the system, and η is a critical exponent. At high temperatures the correlation decays to zero exponentially with the distance between the spins. ξ is the correlation length which is a measure of the range over which fluctuations in one region of space are correlated with or influence those in another region. Two points separated by a distance larger than the correlation length will each have fluctuations that are relatively independent. Experimentally, the correlation length is found to diverge at the critical point which means that distant points become correlated and long-wavelength fluctuations dominate. The system 'loses memory' of its microscopic structure and begins to display new long-range macroscopic correlations. I will say more about this 'memory loss' below but for now let me just point out that, while not the whole story, it is nevertheless a significant

component of the “non-causal” interpretation of RG in that it facilitates the independence between the micro and macro levels in certain types of systems.

An important feature associated with singular behaviour and the variation over a large range of length scales is the way physical quantities follow the changes in scale. In RG calculations these changes result from the multiplication of several small steps to produce a large change in length scale l . As the length scale changes, so do the values of the different parameters describing the system. The change in the parameters is implemented by a beta function

$$\left\{ \tilde{J}_k \right\} = \beta \left(\{ J_k \} \right)$$

which induces what is known as an RG flow on the J -space, the space of couplings. $\{ J_k \}$ is a set of coupling constants where the values of J under the flow are called running couplings which refers to the dependence of a coupling on scale changes. The coupling can refer to any interaction such as the connection between spins, or in the Ising model, the tendency of neighbouring spins to be parallel. Each transformation increases the size of the length scale with the phase transition identified with a fixed point where further iterations of the RG transformations produce no changes in the correlation length (or couplings in QFT). Hence, the fixed points give the possible macroscopic states of the system at a large scale. Although the correlation length diverges at critical point, using the RG equations in effect reduces the degrees of freedom which in turn reduces the correlation length. Fewer degrees of freedom imply new couplings, but no change in the physics. This result incorporates both scale-invariance and universality. The significance of Wilson’s (1975) approach is that you can consider all possible couplings so there is no need to “decide” which ones to focus on, nor speculate what the outcome of the large scale will be. One simply follows the renormalization procedure which will bring you to a fixed point.⁵

Behaviour near critical point is described using power laws where some critical property is written as a power of a quantity that might become very large or small.⁶ The behaviour of the order parameter, the correlation length, and correlation function are all associated with power laws where the “power” refers to the critical exponent or index of the system. Diverse systems (liquid and magnets) exhibit the same scaling behaviour as they approach critical point and take the same values for the critical

⁵ In addition to Wilson’s and Kadanoff’s works there are several comprehensive general discussions of RG in the physics literature some of which include Fisher (1998), Goldenfeld (1993), and Zinn-Justin (2002) as well as Wilson’s own (1983) Nobel lecture and, for a more popular version, his (1979) *Scientific American* article.

⁶ A power law is essentially a functional relationship between two quantities, where one quantity varies as a power of another. Power-law relations are sometimes an indication of particular mechanisms underlying phenomena that serve to connect them with other phenomena that appear unrelated (universality). Some examples of power laws include the Gutenberg–Richter law for earthquakes, Pareto’s law of income distribution, and scaling laws in biological systems.

exponents, which indicates that they belong to the same universality class, a fact that, as we shall see later, can only be explained via RG.

So, why exactly do we need RG to understand what's going on in phase transitions and to explain the foundations of universality? The main problem is that systems near T_c depend on two different length scales, the microscopic scale given by atoms or lattice spacing and the dynamically generated scale given by the correlation length which characterizes macro phenomena. In many classical systems one can simply decouple these different scales and describe the physics by effective macroscopic parameters without reference to the microscopic degrees of freedom. In statistical mechanics this approach became known as mean field theory (MFT) (Landau 1937) and assumes the correlations between stochastic variables at the micro scale could be treated perturbatively with the macro expectation values given by quasi-Gaussian distributions in the spirit of the central limit theorem.⁷

MFT predicted a universality of the singular behaviour of thermodynamic quantities at T_c , meaning that they diverged in exactly the same way; for instance, ξ always diverges as $(T - T_c)^{-\nu}$. It assumed these properties were independent of the dimension of space, the symmetry of the system, and the microphysical dynamics. However, it soon became apparent that experimental and theoretical evidence contradicted MFT (e.g., Onsager's 1944 exact solution to the 2D Ising Model). Instead critical behaviour was found to depend not only on spatial dimensions, but on symmetries and some general features of the models. The fundamental difficulty with MFT stems from the very problem it was designed to treat—criticality. The divergences at T_c were an indication that an infinite number of stochastic degrees of freedom were in some sense relevant to what happens at the macro level, and it was exactly these fluctuations on all length scales that would add up to contradict the predictions of MFT.

The type of behaviour we witness at critical point is unlike the typical case where physical systems have an intrinsic scale or where other relevant scales of the problem are of the same order. In these latter contexts phenomena occurring at different scales are almost completely suppressed with no need for any type of renormalization. Such is the case with planetary motion; it is possible to suppress, to a very good approximation, the existence of other stars and replace the size of the sun and planets by point-like objects. And, in non-relativistic quantum mechanics we can ignore the internal structure of the proton when calculating energy levels for the hydrogen atom. However, in MFT we have exactly the opposite situation; divergences appear when one tries to decouple different length scales. The divergence of ξ makes it impossible to assume a system of size L is homogeneous at any length scale $l \ll L$, and, because ξ also represents the size of the microscopic inhomogeneities in the system its divergence prevents the statistical fluctuations from being treated perturbatively. Hence, the impossibility of using statistical averaging techniques for these types of systems.

⁷ I will have more to say about the relationship between RG methods and the central limit theorem below.

This failure of MFT is interesting from the point of view of generic explanations. Because the statistical averaging procedures cannot accommodate the way inhomogeneities in the microscopic distributions contribute to large scale cooperative behaviour, the task was to explain how short-range physical couplings could generate this type of behaviour at the macro level and how to predict it. What the situation seemed to imply was that it wasn't the values of specific physical quantities that were relevant but rather the features of their dependence with respect to the size N of the system and the control parameters K .⁸ In other words, the way micro features cooperated to produce universal behaviour was the object of explanation, a general feature that could be separated from more specific aspects of the microscopic dynamics. RG methods provided a solution to such problems by determining, in a recursive manner, the effective interactions at a given scale and their relation to those at neighbouring scales. This is one way we can think of RG as providing non-causal information: it illustrates a separation between micro constituents and macro behaviour and does so in a way that is distinct from ordinary statistical averaging procedures where the micro processes remain linked to macro behaviour. And, as I noted in the short discussion of statistical mechanics in section 2, while this latter type of explanation is probabilistic it nevertheless embodies a causal component. However, highlighting this aspect of the non-causal character of RG is not sufficient: further considerations are necessary in order to see the extent to which it functions as an explanatory framework.

4. Structural Features of RG Explanations

We have already briefly seen how the RG methodology differs from MFT. Although it involves a process similar to averaging the ensemble over small scale correlations, there is an important difference—instead of using the ensemble to calculate an average, as in statistical mechanics, we use RG to transform one ensemble into another one with different couplings. Each transformation increases the length scale so that it eventually extends to an infinite limit, with this infinite spatial extension responsible for determining the thermodynamic singularities included in the calculation. But again, the way in which this process takes place is vital for understanding the differences between RG methods and statistical averaging.

In order to fully understand how the fixed points become the focal point for explanation it is important to stress that the basis of the idea of universality is that the fixed points are a property of *transformations* that are not sensitive to the original Hamiltonian. What the fixed points do is determine the kinds of cooperative behaviour that are possible, with each type defining a universality class. The coincidence of the critical indices in very different phenomena was inexplicable prior to RG methods which were successful in showing that the differences were related to irrelevant

⁸ A control parameter is one that appears in the governing equations of a system and measures the effects of an exterior influence such as temperature, pressure, field intensity, etc.

observables—those that are “forgotten” as the scaling process is iterated. But, it isn’t simply the elimination of irrelevant degrees of freedom that is important here, it is the existence of cooperative behaviour characterized by the fixed points that serves as the explanatory foundation of universality. The fact that RG enables us to determine the existence of fixed points suggests that the explanation of universality is, at its foundation, a mathematical one.

What justifies this claim?—especially since the elimination of unwanted degrees of freedom coincides with the suppression of information related to explanation at different levels, a strategy that is common in all areas of physics and is also embedded in the statistical averaging procedures in SM. What is different in the context of RG, and what makes the explanation non-causal, is the *way* information is suppressed and what the end result is. If we simply average over microscopic information then that information still plays a causal role in explaining the outcome; in the case of RG the recursive application of the transformation is not a statistical average but, as noted involves the creation of a new ensemble with different values for the parameters. A significant feature of RG is that it illustrated how, in the long wavelength/large space-scale limit, the scaling process in fact *leads* to a fixed point when the system is at a critical point, with very different microscopic structures giving rise to the same long-range behaviour. But, it is also important to note here that this isn’t an instance of multiple realizability. If it were we could simply appeal to each of the realizers (the different microstructures) as the source of causal information. Instead, the application of RG transformations eliminates this microstructure from consideration leaving the critical behaviour to be explained via the fixed points. And, what is especially crucial for the non-causal account is that the fixed points are defined in a purely mathematical way; a fixed point of a function is simply an element of the function’s domain that is mapped to itself by the function. c is a fixed point of $f(x)$ if and only if $f(c) = c$. Hence $f(f(\dots f(c)\dots)) = f^n(c) = c$, becomes an important terminating consideration when recursively computing f .⁹

An application of RG methods involves transferring the problem from a study of a particular system S to a study of scale transformations such that the results depend only on the scaling properties. What that requires is a shift away from the phase space of the system to a space of Hamiltonians. This space of Hamiltonians is sometimes referred to as the space of couplings that I mentioned earlier where the Hamiltonian is the function that acts on the coupling constants J . The transformations take place within this functional space with each element corresponding to a physical system with some fixed value of the control parameter(s) K . It is important to note that as the scale changes the general form of the Hamiltonian also changes so that the renormalized Hamiltonian will take on a more or less generic, mesoscopic form (Fisher 1998). Rather than study the equilibrium state of S within a specified model (computing the

⁹ From the perspective of phase transitions you can have linearization around an unstable fixed point which gives the appearance of phase changing behaviour, but a definite phase change requires stable fixed points which one only gets under the assumption of an infinite system.

value of state functions and variations with respect to variables and parameters) the focus is on the *transformation* of the model and its parameters in connection with a change in scale of the description of *S*. This allows for the calculation of quantitative and universal results from properties of the renormalization flow.

What this means is that RG equations show that critical point phenomena have an underlying order. Indeed what makes the behaviour of these phenomena predictable, even in a limited way, is the existence of certain scaling properties that exhibit universal behaviour. The number and type of relevant parameters is determined by the *outcome* of the renormalization calculation.¹⁰ Assuming that a fixed point is reached one can find the value that defines the critical temperature and the series expansions near the critical point provide the values of the critical indices. The nontrivial fixed points that represent the critical states are such that each distinct Hamiltonian whose trajectory converges to the same fixed point will be identical with respect to the nature of their criticality and the free energy in their neighbourhood. In that sense RG methods provide us with both mathematical and physical information concerning how and why different systems exhibit the same behaviour near critical point. They determine these universality classes by proving the existence and universality of scaling laws, laws that provide the mathematical foundation for observed experimental behaviour.

This kind of explanation differs from Lange's account in that the framework doesn't require us to determine whether, in the context, the explanation is mathematical; nor is there any room for assessing the degree to which the mathematics (as opposed to the physics) is the primary explanatory vehicle. The only question is whether one accepts RG as an explanatory framework (rather than just a calculational technique). But, as we saw earlier, the reasons for characterizing it as explanatory centre largely on the role of the fixed points. As a product of the RG transformations the fixed points ground the explanation of universality, and like the transformations themselves, are purely mathematical objects or the outcome of a purely mathematical process. Moreover, unlike Lange's double pendulum, there is no accompanying causal story that one can appeal to as a way of "understanding" or reinterpreting the mathematical framework. To put the point in slightly stronger terms: without the mathematics of RG the physical phenomenon of universality is simply a mystery. Of course this in no way undermines Lange's argument; instead my claim is that RG offers us a case that is independent of context and degree in its status as a purely mathematical, non-causal explanation.

The next step in the explanatory story is to show how, in more detail, these scaling (mathematical) properties deliver *physical* information about the systems we are interested in, as well as the structural, non-causal nature of that information. Spelling this out requires that we further differentiate RG explanations from probabilistic ones since the latter can easily be assimilated into a causal framework.

¹⁰ In earlier versions parameters like mass, charge, etc. were specified at the beginning and the change in length scale simply changed the values from the bare values appearing in the basic Hamiltonian to renormalized values. The old renormalization theory was a technique used to rid quantum electrodynamics of divergences but involved no "physics".

5. RG and Probabilistic Explanations: Disentangling the Differences

From this very brief description we can see the shift away from methods used in statistical mechanics that are grounded in probability theory (and possibly causality) toward a more structurally based approach to dealing with large scale features of phenomena. However, there are also similarities, similarities that need to be carefully spelled out so that the differences can be properly understood. Khinchin (1949) was responsible for the first systematic use of probabilistic methods in statistical mechanics and showed, using the central limit theorem (CLT), that the Boltzmann distribution of the single-molecule energy in systems of weakly correlated molecules is universal. In other words, it is independent of the form of the interaction provided it is short range (Jona-Lasinio 2001). Although the CLT has a number of variants, in very general terms it states that the sum of many independent random variables tends to a Gaussian, whatever the original distribution might have looked like. When we sum many random variables the details of the distributions of the individual variables becomes unimportant—instead, simple generic behaviour emerges. Put differently, the Gaussian distribution is a fixed point function for large sums.¹¹ As we have seen earlier, failures of MFT could be traced to the existence of strongly correlated variables, cases to which the central limit theorem fails to apply. Yet, we can derive the CLT from within the framework of RG, so what exactly is the connection between them?

Without going through the specific steps (see Sornette 2000) the RG derivation of CLT shows that the Gaussian distribution is a fixed point in the sense that it is form-stable under successive renormalizations. This is essentially the notion of self-similarity that is crucial for scale invariance, which means that the resulting Gaussian function is identical to the initial Gaussian function after the appropriate shift and rescaling of the variables. This point was clarified by Cassandra and Jona-Lasinio (1978) who showed that scaling is strictly connected with the generalizations of limit theorems. In the original articles on RG by Wilson (1975) and others, no simple probabilistic interpretation was given but what Cassandra and Jona-Lasinio argue is that the universality of critical behaviour is connected with the universality of limit distributions obeying certain constraints (1978: 919), specifically a fixed point equation. It is important to note, however, that CLT (especially the RG derivation) is the expression of a collective phenomenon that holds, strictly speaking, in the limit of infinite N (although the Gaussian shape is a good approximation of the centre of the probability distribution function for the sum if N is sufficiently large) and hence says nothing about behaviour of the tails for finite N .

¹¹ The idea is nicely expressed by Gnedenko and Kolmogorov (1954: 1) who claim that “all epistemological value of the theory of probability is based on this: that large scale random phenomena in their collective action create strict, non-random regularity.”

We know that collective behaviour in critical phenomena involves dependent variables; hence, clarifying the relation between CLT and RG involves generalizing CLT to the case of dependent variables where the random fields appearing in these new limit theorems for critical phenomena will display the necessary scaling properties. The problem is that no general account of these limit theorems is presently available. Some examples have been given by Jona-Lasinio (2001) and others but these are far from satisfactory since a form of scaling is introduced from the beginning. What this difference highlights is the gap between the probabilistic aspects we normally attribute to systems governed by statistical mechanics and the way we need to understand critical behaviour as governed by RG methods. The latter highlights the deep statistical nature of critical universality, specifically the importance of distinguishing between physical couplings at the micro scale and the statistical correlations between these couplings. In other words, the essential feature in explaining the behaviour of critical phenomena is not the physical couplings themselves but the statistical coupling of these couplings (Lesne 1998). The RG flow takes place in this space of Hamiltonians (or coupling constants) and illustrates the statistical self-similarity between the system at critical point and the original.

As Lesne (1998) points out, we can see how these are importantly different by considering the range λ of the direct physical couplings between elementary constituents compared with the range ξ of statistical correlations. In the Ising model λ , the lattice parameter, is the distance a separating the nearest neighbour spins on the lattice. It is a physical quantity that depends on the value and doesn't determine the critical behaviour of the system. By contrast, ξ diverges at T_c in the Ising model; it is always greater than λ and is a global statistical characteristic. In other words, a statistical organization of the constituents and their statistical couplings is present on every scale from λ to ξ . More importantly, any elementary subsystems separated by any length l between λ and ξ will have correlated statistical distributions whether or not they have any physical interaction. What this means is that the only characteristic scale of the global behaviour of the system is the correlation length ξ , λ (the microscopic scale) plays no role. When a large number of degrees of freedom exhibit collective, cooperative behaviour, it is the statistical characteristics that determine the corresponding macroscopic critical behaviour.

Despite this emphasis on the statistical nature of RG transformations, any truly probabilistic interpretation will require limit theorems for the strongly correlated variables. Because these are not available, our understanding of the process relies on the existence of fixed points and the replacement of one ensemble description with another via a "structural" transformation rather than a straightforward averaging. As I noted earlier, in the case of statistical mechanics the statistical ensemble defined by a Hamiltonian is used to calculate an average but with an RG operation we move from one ensemble to another with a Hamiltonian that has different couplings from the original one. The mathematical foundation of these types of structural transformations comes from dynamical systems theory which deals with changes under

transformations and flows toward fixed points. Instead of deriving exact single solutions for a particular model the emphasis is on the geometrical and topological structure of ensembles of solutions.

It is tempting to think of the Hamiltonians in RG transformations as somehow encoding the details of the micro-level components and correspondingly explaining universality in terms of the component behaviour and the flow to fixed points. This picture seems to commit us to the micro-components as having a role to play in the explanation even though they may be washed out in the RG transformation. As Reutlinger (2017: 2303), a proponent of this view, remarks: “the system having a *particular micro-structure* S (represented by a Hamiltonian) determines the fact that this kind of system belongs to a universality class U ” (my emphasis). He claims that facts about the interacting components ‘fix’ the universality class to which the system in question belongs. And, if two physical systems belong to different universality classes, then the systems differ, for instance, with respect to the spatial dimension of the system and on the symmetry properties of the order parameter. A difference, for instance, in spatial dimensionality will be accompanied by a difference on the level of the *components*.

While it is true that spatial dimension and symmetry of the order parameter are important features in determining universality classes, it is also important to distinguish these constraints from the “components” one identifies with microstructure. Spatial dimensionality is independent of microphysical properties except in the sense that systems with different properties can have the same dimensionality; but the latter has no bearing on the former. Similarly, the particular symmetry associated with the order parameter is distinct from particular features of the microphysics, unlike the gauge symmetries I mentioned at the beginning where local gauge invariance can determine the form of the field interactions. To refer to symmetry and dimensionality as “components” is to equate the structural constraints on a system with its material constituents giving us a distorted picture of the micro–macro relations in RG explanations. It is only the vectorial or tensorial character of the relevant order parameter (e.g., scalar, complex number *alias* two-component vector, three component vector, etc.) and dimensionality that are crucial for defining the universality class in terms of the values for critical exponents; the lattice structure is irrelevant. For instance, the excluded-volume problem for polymers was known to have closely related but *distinct* critical exponents from the Ising model, whose values depended on dimensionality but *not* lattice structure.

Of further importance here is the fact that the RG Hamiltonians are not the ordinary phase space Hamiltonians we write down when dealing with physical systems. Normally in condensed matter physics one focuses on some specific form of H with at most two or three variable parameters—the Ising model is a simple example with just two variables, t , the reduced temperature, and h , the reduced field. An important feature of Wilson’s approach, however, is to regard any such “physical Hamiltonian” as merely specifying a subspace (spanned, say, by “coordinates” t and h) in a very large space of possible (reduced) Hamiltonians. The important point here is to clarify what

the physical import of this picture is. Different symmetries and spatial dimensions produce different fixed points which encode different behaviour identified with different universality classes. And, each universality class shows a connection between an internal symmetry (e.g., Ising model's up-and-down, or rotation in a plane) and the topological properties of a system that extend over an effectively infinite region of space, a region much larger than the range of forces. It shows thermodynamic singularities, correlation functions that fall off algebraically, and internal parameters such as coherence or correlation length.

But again, these aren't "components" of the system in the usual sense, the sense relevant for writing down a phase space Hamiltonian. The coarse-grained patterns one gets for, say, fluids and magnets, don't match the models of these system—the latter are two-dimensional (planar) while the former are typically three-dimensional. As I noted earlier, the spatial dimension d is one of the features not washed out by RG transformations and hence is reflected in the values of the critical indices. The other feature which defines the universality class is the number of components n of the order parameter. The order parameter may take the form of a complex number, a vector, or even a tensor, the magnitude of which goes to zero at the phase transition. In many cases the order parameter is a scalar (e.g., density difference for a fluid). For example, in the case of superfluid He^4 we have an order parameter with two components, the amplitude and phase of the wave function describing the condensate. These systems fall into the same universality class as the XY model where the components of the order parameter $n = 2$ are classical spin vectors. Other features characteristic of a specific universality class such as symmetry breaking perturbations may be relevant or irrelevant but will ultimately depend on d and n . There are, of course, many different universality classes corresponding to different dimensionalities and to different symmetries of the order parameter as in the case of the Ising model, XY-model, and Heisenberg model, respectively having one, two, and three components in their spin vectors. As a result each has different critical fixed points in three dimensions.

The final piece of the puzzle is to spell out in a bit more detail what I mean by the "generic structural" features of RG explanations. By focusing on large scale structural behaviour we can hopefully see how RG techniques furnish an understanding of complex behaviour that extends beyond calculating values for critical indices in phase transitions.

6. Dynamical Systems and RG Explanations

In general, explanations in the context of dynamical systems have the following goals: to describe the typical behaviour of trajectories for many different types of systems as time (rather than space) tends to infinity, and to understand how this behaviour changes under perturbations. What we want to know is the extent to which the system is stable; whether it is possible to deform a perturbed system in a way such that we can recover the original one. In that sense the goal is to answer general questions about, for example, whether there is any relation between the long-term behaviour and initial

conditions, rather than trying to find precise solutions to the equations defining the system itself, something that is often not possible. So, a simple definition of a dynamical system can be given in terms of a group of transformations on a topological space (manifold) with one parameter—time. An equation of evolution is then used to generate trajectories extended in time. The objective is to describe the fixed points—values of the variable(s) describing the steady states of the system that won't change over time. If a fixed point is attractive, nearby states will converge toward it. In addition to fixed points there are also periodic points, states of the system which repeat themselves after several time-steps. These two features are crucial since simple nonlinear dynamical systems often exhibit chaotic behaviour that is more or less random and unpredictable.

Feigenbaum (1978) was responsible for showing that a class of dynamical systems could exhibit universal self-similar behaviour that could be explained using RG. The difference with critical phenomena is that the spatial extension of the system is replaced by duration in time and R , the renormalization operator, acts on the evolution law for the system. Feigenbaum's focus was on the logistic map which is one of the simplest forms of a chaotic process. As with any one-dimensional map, it is a rule for getting a number from a number. Mathematically the logistic map is written

$$x_{n+1} = rx_n(1 - x_n) \quad (n = 0, 1, 2, \dots)$$

where x_n is a number between 0 and 1 and r is a parameter that can be varied to change the character of the x sequence. The logistic map has been used extensively to study variability in populations of different species (May 1976). The simplest behaviour of an orbit of such a map is convergence to a fixed point. Some mathematical mappings involving a single linear parameter exhibit apparently random behaviour (chaos) when the parameter lies within certain ranges. As the parameter is increased toward this region, the mapping undergoes bifurcations at precise values of the parameter. So, if r (the driving parameter) is small, the values of the iteration eventually just repeat (system becomes periodic). As r increases (3.0), there is oscillation between two values, a bifurcation results and the periodicity doubles.¹² As r increases, there is further bifurcation with oscillation between four values and so on. Chaos ensues when r equals 3.57. Feigenbaum discovered that the ratio of the difference between the values at which such successive period-doubling bifurcations occur tends to a constant of around 4.6692... (a ratio of convergence known as the Feigenbaum constant). Not only did he provide a mathematical proof of that fact but he showed that the same behaviour, with the same mathematical constant, would occur within a wide class of mathematical functions, prior to the onset of chaos.

What this means is that in the limit of a-periodic behaviour, there is a unique and hence universal solution common to all the different systems undergoing period

¹² With r between 0 and 1, the population will eventually die, independent of the initial population. With r between 1 and 2, the population will quickly approach the value $r-1/r$ independent of the initial population. With r between 2 and 3, the population will also eventually approach the same value but will first fluctuate around that value for some time.

doubling. They are analogous to critical phenomena in that they are fixed-point theories, with the Feigenbaum constant δ (the rate of convergence/onset of complex behaviour) viewed as a critical exponent. In other words, for all systems undergoing period doubling, δ has a *universal* value. In that sense the transition to chaos can be seen as a critical phase transition and treated using RG methods. In the application to dynamical systems this involves proving the scaling behaviour of the bifurcation values; predicting the numerical value of δ and describing the associated universality class.

The first step is to define the relevant space F of evolution maps on which the renormalization operator will act. This is directly analogous to the space of Hamiltonians for the spatial (position) case. Once the fixed points $R\varphi = \varphi$ are found it is possible to show that the fixed point equation

$$[\varphi \circ \varphi] [x\varphi^2(0)] = \varphi^2(0) \varphi(x) \quad \varphi \in F$$

where $\varphi(0) = 1$ admits a unique solution in F . The equation expresses the exact self-similarity between φ and its iterate $\varphi \circ \varphi$ at all the time scales between the trajectories that they generate. One then investigates the linear stability of φ with respect to the renormalization action in order to determine the flow generated by R in F . The renormalization picture in the space F of unimodal maps is then related to the period doubling scenario observed in most of the one-parameter families of such maps. Without going through the details of each of the results the outcome is that the universality classes of the period doubling scenario is the set of all one-parameter families of unimodal maps that cross transversally the basin of attraction of φ with respect to the renormalization action.¹³

The renormalization procedures in each of these cases—statistical mechanics and dynamical systems—illustrate the similarities between them.¹⁴ But, what is more important than the analogies is the underlying structure that makes them possible. Each context, the Ising model in statistical mechanics and different types of dynamical systems, has a specific structural law or constraint that embodies the information necessary to describe an equilibrium state or its evolution. In statistical mechanics the relevant structural feature is the Hamiltonian while in dynamical systems it is the evolution map. RG methods focus on the *space* defined in terms of the Hamiltonians or the evolution maps or whatever the relevant structural feature for the system is. Shifting away from individual “physical” models of the system with a specific micro-structure defined on a phase space and replacing it with an emphasis on the way the structural features (the space of Hamiltonians or evolutionary maps) change under RG transformations illustrates how we can understand the relation between stable macro behaviour and the micro level from which it arises.

This change of orientation/methodology results in a rather different epistemological situation. Emphasis on a model of a system S fails to provide a way of focusing on the

¹³ See Feigenbaum (1978).

¹⁴ See Lesne (1998).

right degrees of freedom for the problem at hand. In contrast, the emphasis on renormalization flow/maps allows us to investigate the robustness of predictions via their reliance on structurally stable behaviour. As Goldenfeld and Kadanoff (1999) point out, complexity can be defined as structure with variations. They point out that nature can produce complex structures even in simple situations and can obey simple laws in complex situations. This is exactly what RG so powerfully illustrates! Explaining the structural stability of complex systems in terms of structural constraints and how they transform might sound like an obvious strategy but the processes involved in carrying it out were far from obvious until the advent of RG methods.

7. Conclusions

One of the mainstays of my argument has been the importance of structure rules in characterizing RG methods as a form of mathematical explanation. This is accomplished by showing how the source of non-causal information about the system's evolution and stability comes via the transformation of structural features of systems (the Hamiltonian in SM, the evolution map for discrete dynamical systems, etc.) rather than specific values for microscopic parameters. The only signature of short distance microscopic behaviour lies in the initial conditions of the RG flow, not in the flow itself. Ultimately the objective is the determination of fixed points which provide the basis for the structural stability characteristic of cooperative behaviour.

The typical approach in theoretical physics is to consider a given model and attempt to extract information by studying its evolution, equilibrium state, and the solutions. However, there is often no explicit procedure for incorporating idealizations and approximations or for determining which scales and degrees of freedom are important. What the RG framework does is show in very explicit ways the relation between certain features of macro behaviour and their relation to changes in scale. Instead of investigating a specific model, focusing on RG flows allows us to investigate structural stability and provide robust predictions by transforming qualitative information (the belonging to the same universality class) into quantitative information (the values of the critical exponents and expression of scaling functions).

Emphasizing the way RG analysis proceeds via these structural features, drawing out similarities between the use of RG in SM and dynamical systems, as well as highlighting its connections with the central limit theorem enable us to appreciate the power of RG methods in investigating properties of a wide variety of physical systems.

Acknowledgements

Support of research by the Social Sciences and Humanities Research Council of Canada and the Alexander von Humboldt Foundation is gratefully acknowledged. I would also like to thank the editors for their helpful comments and suggestions.

References

- Baker, A. (2005), 'Are there Genuine Mathematical Explanations of Physical Phenomena?', *Mind* 114: 223–38.
- Baker, A. (2009), 'Mathematical Explanation in Science', *British Journal for the Philosophy of Science* 60: 611–33.
- Bangu, S. (2012), *The Applicability of Mathematics in Science: Indispensability and Ontology* (London: Palgrave Macmillan).
- Batterman, R. (2010), 'Reduction and Renormalization', in A. Hüttemann and G. Ernst (eds.), *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics* (Cambridge: Cambridge University Press), 159–79.
- Cassandro, M. and Jona-Lasinio, G. (1978), 'Critical Point Behaviour and Probability Theory', *Advances in Physics* 27: 913–41.
- Feigenbaum, M. (1978), 'Quantitative Universality for a Class of Nonlinear Transformations', *Journal of Statistical Physics* 19: 25–52.
- Fisher, M. E. (1998), 'Renormalization Group Theory: Its Basis and Formulation in Statistical Physics', *Reviews of Modern Physics* 70: 653–81.
- Gell-Mann, M. and Low, F. E. (1954), 'Quantum Electrodynamics at Small Distances', *Physical Review* 95: 1300–12.
- Gnedenko, B. V. and Kolmogorov, A. N. (1954), *Limit Distributions for Sum of Independent Random Variables* (Reading, MA: Addison Wesley).
- Goldenfeld, N. (1993), *Lectures on Phase Transitions and the Renormalization Group* (Reading, MA: Addison-Wesley).
- Goldenfeld, N. and Kadanoff, L. (1999), 'Simple Lessons from Complexity', *Science* 284: 87–9.
- Jona-Lasinio, G. (2001), 'Renormalization Group and Probability Theory', *Physics Reports* 352: 439–58.
- Kadanoff, L. (1966), 'Scaling Laws for Ising Models near T_c ', *Physics* 2: 263–72.
- Kinchin, A. I. (1949), *Mathematical Foundations of Statistical Mechanics* (New York: Dover).
- Landau, L. D. (1937), 'On the Theory of Phase Transitions'. Translated and reprinted from L. D. Landau, *Collected Papers*, vol. I (Moscow: Nauka, 1969), 234–52.
- Lange, M. (2013), 'What Makes a Scientific Explanation Distinctly Mathematical?', *British Journal for the Philosophy of Science* 64: 485–511.
- Lesne, A. (1998), *Renormalization Methods: Critical Phenomena, Chaos, Fractal Structures* (New York: Wiley).
- May, R. M. (1976), 'Simple Mathematical Models with Very Complicated Dynamics', *Nature* 261: 459–67.
- Onsager, L. (1944), 'Crystal Statistics. I. A Two-Dimensional Model with an Order-Disorder Transition', *Physical Review* 2: 117–49.
- Pincock, C. (2007), 'A Role for Mathematics in the Physical Sciences', *Noûs* 41: 253–75.
- Pincock, C. (2012), *Mathematics and Scientific Representation* (New York: Oxford University Press).
- Reutlinger, A. (2014), 'Why Is There Universal Macro-Behavior? Renormalization Group Explanation as Non-Causal Explanation', *Philosophy of Science* 81: 1157–70.
- Reutlinger, A. (2017), 'Are Causal Facts Really Explanatorily Emergent? Ladyman and Ross on Higher-Level Causal Facts and Renormalization Group Explanation', *Synthese* 194: 2291–305.

- Sklar, L. (1993), *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics* (Cambridge: Cambridge University Press).
- Sornette, D. (2000), *Critical Phenomena in the Natural Science* (Dordrecht: Springer).
- Steiner, M. (1978), 'Mathematical Explanation', *Philosophical Studies* 34: 135–51.
- Wilson, K. (1971), 'The Renormalization Group (RG) and Critical Phenomena 1', *Physical Review B* 4: 3174–83.
- Wilson, K. (1975), 'The Renormalization Group: Critical Phenomena and the Kondo Problem', *Reviews of Modern Physics* 47: 773–839.
- Wilson, K. (1979), 'Problems in Physics with Many Scales of Length', *Scientific American* 241: 158–79.
- Wilson, K. (1983), 'The Renormalization Group and Critical Phenomena', *Reviews of Modern Physics* 55: 583–600.
- Zinn-Justin, J. (2002), *Quantum Field Theory and Critical Phenomena* (Oxford: Clarendon Press).

PART III

Beyond the Sciences

Two Flavours of Mathematical Explanation

Mark Colyvan, John Cusbert, and Kelvin McQueen

1. Introduction

Explanation in mathematics is puzzling. Mathematicians tell us that some proofs are explanatory while others are not.¹ That is, all proofs establish the theorem in question but some proofs go further and explain why the theorem holds.² But what kind of thing is an explanatory proof? Some of the usual candidates for explanation in science do not seem to work for mathematics. For example, some take explanation to be closely related to causal history but there is no place for causation in mathematics. Similar difficulties arise for counterfactual and interventionist accounts of explanation; mathematics, if true, is a body of necessary truths, so there does not seem to be any room for counterfactuals or intervening.³

If we focus on proofs as the locus of explanation in mathematics,⁴ one rather natural thought is that mathematical explanations have something to do with the structure of the proof—the explanatory proofs have some especially desirable structure that reveals the reason for the theorem holding.⁵ Although we will not argue against this view here,⁶ we find it implausible that explanation can be characterized entirely in terms of

¹ For example, see Gowers and Neilson (2009: 879).

² Although we occasionally use the less clumsy realist language of mathematical “truths” and “facts”, in this chapter we wish to sidestep realism–anti-realism issues. If you’re a mathematical realist, explanatory proofs tell us why the theorem is *true*. If you’re a mathematical anti-realist you may not believe that the theorem in question is true. You might, instead, think that the theorem is “true-in-the-fiction of mathematics” or some such. In any case, you can, and should, still countenance the distinction between explanatory proofs and non-explanatory ones. The former, may, for example, provide an intra-fiction explanation of the fictional result, just as there are explanations in literary fiction of why some fictional character behaved as she did.

³ Although see Baron et al. (2017) for some moves in this direction.

⁴ It’s not clear that proofs are the only place where explanation arises. For example, it might be argued that we find explanation in domain extensions (Colyvan 2012: ch. 5).

⁵ See for example an exchange between Alan Baker (2010) and Marc Lange (2009) on the explanatoriness of proofs by mathematical induction.

⁶ See Colyvan (2012: ch. 5) for such an argument. For example, in some cases *reductio* proofs can be transformed into constructive proofs and, in such cases, it seems implausible that the former are not

the structure of the proof. In any case, in this chapter we will dig a little deeper—below the level of the structure of the proofs.

To be clear about our target, it's worth distinguishing the kind of explanation we're interested in here from another that's prominent in the literature. *Intra-mathematical explanation* is the explanation of one mathematical fact in terms other mathematical facts. This is to be contrasted with *extra-mathematical explanation*, which is the explanation of some physical phenomenon via appeal to mathematical facts. The existence of such extra-mathematical explanation is still somewhat controversial.⁷ We will be firmly focused on intra-mathematical explanation. More specifically, our interest is in the intra-mathematical explanation found in proofs of theorems.⁸

We will look, in some detail, at two different proofs of an important result in group theory: the Free Group Theorem. Each of these two proofs has some claim to being an explanatory proof. We explore whether these proofs share a common feature that accounts for their explanatoriness. We conclude that the two proofs exhibit two quite different explanatory virtues. We make cases for two plausible, but competing, accounts of mathematical explanation and we suggest that there might be more than one kind of explanation at work in mathematics.

2. A Few Words about Methodology

Neither the Free Group Theorem nor the two proofs we offer are straightforward but we make no apology for this. It is, in our view, important to tackle examples from more advanced mathematics. It would be all too easy to be misled by focusing on elementary examples from high-school mathematics. What is required is a systematic study of proofs from various areas of contemporary mathematics—analysis, abstract algebra, topology, number theory, and so on. The examples also need to go beyond high-school mathematics.⁹ Of course, there will be limits to how advanced the mathematics can be in order for philosophers of mathematics—who are, after all, typically not professional mathematicians—to understand it and be able to draw reliable philosophical morals.¹⁰ Still, those wishing to take our word on the technical

explanatory while the latter are. In such cases, either they are both explanatory or neither are. Either way, there's more to it than merely the structure of the proof.

⁷ See Baker (2005, 2009), Baron (2014), Baron and Colyvan (2016), Colyvan (2001, 2002, 2010), and Lyon and Colyvan (2008) for examples of extra-mathematical explanations.

⁸ In the past this has received less attention in the philosophical literature on explanation (Resnik and Kushner 1987; Steiner 1978a, 1978b) although that seems to be changing, with a number of recent contributions to this topic (Colyvan 2012 and forthcoming; Giaquinto 2016; Hafner and Mancosu 2005; Lange 2014, 2016; Mancosu 2008a, 2008b; Pincock 2015; Raman-Sundström and Öhman forthcoming).

⁹ Marc Lange has already started this project. In his paper (Lange 2014) he discusses some more advanced examples. The present chapter can be seen as another step in that direction, although the conclusions we draw from our example are not the same as Lange's conclusions.

¹⁰ Moreover getting on top of proofs from several different areas of contemporary mathematics can be challenging, even for professional mathematicians.

details of the proofs and our interpretations of them can skip to the discussion for the philosophical upshot.

Ideally, we need the judgements of mathematicians on which proofs are, and which are not, explanatory. But mathematicians are notorious for covering their tracks in their written work and rarely commit to print judgements of the explanatory powers of proofs. But as anyone who has spent time with mathematicians knows, such judgements are forthcoming in the tea room, in the pub, and even in the classroom. In order to get started on this project we need to scour the literature for the few places where mathematicians do offer judgements on whether the proofs in question are explanatory.¹¹ Beyond this, talking to, or formally surveying, mathematicians are the obvious ways forward. We decided to informally survey mathematicians on discussion forums, where some, at least, are inclined to give their opinions about such matters.¹² The forum discussion led to our investigation of the Free Group Theorem, in part because the mathematical community seemed to be divided on which proofs of this theorem are explanatory. It's often more fruitful to start with easy cases, but we were intrigued by this theorem and the dispute over its proofs.¹³

To anticipate our conclusions and help see where we are heading with the proofs and subsequent discussion, we suggest that the two proofs in question have different and competing claims for explanatory virtue. The first proof—the so-called constructive proof¹⁴—delivers the theorem in question via a detailed construction of the group in question and can be thought to be aligned with a model of reductive explanation in science. The second proof—the abstract proof—delivers the theorem by showing how it is one of a more general class of such theorems and as such, this proof can be thought to be aligned with a unificatory model of explanation. Indeed, the fact that this theorem has two such proofs is one of the reasons we chose to focus on it as our case study.¹⁵

Another reason for focusing on the Free Group Theorem is that it is an important result; it is a central result in group theory, especially with respect to the presentation of groups, but it is also important for other, related areas of mathematics (e.g., hyperbolic geometry). Moreover, the result and the proofs we discuss are interesting in their own right. Enough about methodology, let's get into the mathematics.

¹¹ For example: Aigner and Ziegler (2010), Davis and Hersh (1981), and Hardy (1967).

¹² See Ingliss and Aberdeen (2015) for some interesting formal survey-based work getting at mathematicians' judgements about the virtues of various mathematical proofs.

¹³ We intend to follow up this present chapter with further examples to see if our rather speculative conclusions hold up elsewhere in mathematics.

¹⁴ This name is not meant to suggest that the proof is intuitionistically valid; "constructive" is being used in the non-technical sense here.

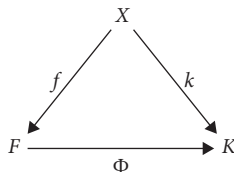
¹⁵ It is important to note that the salient difference between the two proofs is not simply that the abstract proof delivers mere existence whereas the constructive proof constructs an example. There are several interesting differences between the two proofs and this is why we run through the proofs in some detail. We do not wish to give a superficial gloss on the two proofs but the differences highlighted in the main text of this paragraph do strike us as central.

3. The Free Group Theorem

First, recall the definition of a group: A *group* $(G, \cdot)$ is a set of elements G together with a binary operation that together satisfy the four fundamental properties of closure, associativity, the identity property, and the inverse property.

1. *Closure*: If a and $b \in G$, then $a \cdot b$ is also in G .
2. *Associativity*: The group operation is associative, i.e., for all $a, b, c \in G$, $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
3. *Identity*: There is an identity element $e \in G$ such that $e \cdot a = a \cdot e = a$, for every $a \in G$.
4. *Inverse*: There must be an inverse of each element: for each element $a \in G$, the set contains an element a^{-1} such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.

Definition of Free Group: Let X be a set. Group F is free on X if there is a map $f: X \rightarrow F$ and for any group K and map $k: X \rightarrow K$ there is a unique group homomorphism $\Phi: F \rightarrow K$ such that $k = \Phi \circ f$, that is, so that the following diagram commutes.



(This is sometimes expressed in terms of a *universal property*, where the property in question is that which characterizes free groups up to isomorphism. It is the property of being such that the above diagram commutes.)

The *Free Group Theorem* asserts the existence of free groups. More formally, it states that for any set A , there exists a free group on A . (Or equivalently: Given a set A , there exists a free group with basis A .)

4. Free Group Theorem: The “Constructive” Proof

Here we sketch a constructive proof of this result.¹⁶ The proof has two phases: we first use A to construct a group, and define a function from A to that group; and we then prove that together this group and function form a free group on A . (In fact the second phase involves further construction, as we’ll see.)

Given an arbitrary set A , we first use A to construct a group. The members of our group will be certain kinds of “words”, whose “letters” are built up from members of A . We define the *alphabet* on A as the product $A \times \{-1, 1\}$. Each *letter* in our alphabet is

¹⁶ Our proof sketch relies heavily on Rotman (1965: 343–5), where further details can be found.

thus an ordered pair $\langle a, \epsilon \rangle$ where $a \in A$ and $\epsilon = \pm 1$. For convenience, we abbreviate $\langle a, 1 \rangle$ as a and $\langle a, -1 \rangle$ as a^{-1} , and we call a and a^{-1} *rivals*.¹⁷ (At this stage we avoid the term ‘inverses’ since it presupposes a group operation, surrounding which there will be some complications.) (Example: if our set A is $\{a, b\}$, then the alphabet over A is $\{a, b, a^{-1}, b^{-1}\}$.) We then define a *word* on A as a string of alphabet letters of finite length. (Example: $aba^{-1}b$ is a word of length 4.) The *empty word*, written 1, has no letters and length zero. We define the *rival* of a word w , written w^{-1} , as the word obtained by taking the rival of each letter of w and reversing their order. (Example: the rival of aab^{-1} is $ba^{-1}a^{-1}$.) The *concatenation* of words v and w is written vw . This is the word obtained by affixing the head of w to the tail of v . (Example: if $v = ab$ and $w = ab^{-1}$ then $vw = abab^{-1}$.)

Now, it would be nice if we could take our group on A to be the set of words on A , equipped with the operation of concatenation. But this won’t work. While concatenation is an associative binary operation on words, and the empty word 1 will serve nicely as an identity element, the problem lies with inverses: nonempty words have no inverses under concatenation. (Example: there is no word that yields 1 when concatenated with ab .) The set of words on A is not a group under concatenation.

To address this problem, we define a special class of words. Call a word *reduced* if it contains no adjacent rival letters. (Example: aba^{-1} is reduced but $a^{-1}ab$ is not.) Note that the empty word 1 is reduced. The set of reduced words on A , written W , will be the base set of our group.¹⁸

Again though, things are not as straightforward as we’d like. To make W into a group, we’ll need to specify a binary operation on W . But concatenation is not a binary operation on W , because the concatenation of two reduced words need not be reduced. (Example: ab and $b^{-1}a$.)

Consequently, we define a second binary operation on W , called *juxtaposition* and written $*$, as follows. Let $v, w \in W$ be reduced words. Let u be the longest tail of v whose rival u^{-1} is a head of w . (There’s always some such u : even in the case where vw is reduced, we have $u = 1$.) It follows that there exists a head v' of v and a tail w' of w such that $vw = v'uu^{-1}w'$. (Furthermore, we know that u , u^{-1} , v' , and w' are all reduced, because v and w are.) Deleting central rivals gives us $v'w'$, which is *guaranteed to be reduced*. (If it weren’t, then u wouldn’t have been the longest tail of v such that u^{-1} is a head of u : we could have extended u by at least one letter.) We thus have the *Sandwich Lemma*: for any reduced words $v, w \in W$, there exist reduced words u, v' , and w' such that (i) $v = v'u$ (ii) $w = u^{-1}w'$ and (iii) $v'w'$ is reduced. This allows us to define the juxtaposition of v and w by $v * w = v'w'$. Intuitively, juxtaposition amounts to concatenation with cancelling of central rivals. (Example: if $v = aab$ and $w = b^{-1}a^{-1}bb$,

¹⁷ These abbreviations assume that we don’t already have $a, a^{-1} \in A$. If we did, then we’d have distinct letters $\langle a^{-1}, 1 \rangle$ and $\langle a, -1 \rangle$ both abbreviated as a^{-1} . In this unfortunate case we can either choose an alternative notation for $\langle a, -1 \rangle$ (perhaps a') or maintain the ordered pair notation.

¹⁸ In general, a base set is a kind of building block. Here we mean that W will be the set from which we are able to build the group in question.

then $u = ab$ and $u^{-1} = b^{-1}a^{-1}$, so we have $v' = a$ and $w' = bb$, and so $v * w = abb$.) Juxtaposition (unlike concatenation) is a binary operation on W , since a juxtaposition of reduced words is always reduced.

We've now constructed our putative group on A : the set of reduced words on A , equipped with concatenation, or $(W, *)$. (We're yet to *prove* that it's a group: we'll get to that shortly.) Next, we construct our function from A to W . Define $f : A \rightarrow W$ such that $f(a) = a$ for all $a \in A$. Thus f simply maps each letter in A to its corresponding one-letter word in W .

Now for the second phase of the proof: showing that $(W, *)$ and f form a free group on A . One might naturally begin by showing that $(W, *)$ is a group. And indeed this seems within reach. First, we've seen that juxtaposition is a binary operation on W . Second, since $w * 1 = w = 1 * w$ for all $w \in W$, the empty word serves an identity element. And third (by contrast with concatenation), each $w \in W$ has an inverse in W under juxtaposition, namely its rival w^{-1} : it's easy to show that for each $w \in W$ we have $w * w^{-1} = 1 = w^{-1} * w$. (We can henceforth dispense with talk of rivals and safely speak of w^{-1} as the inverse of w in W .)

Unfortunately (again by contrast with concatenation), proving the associativity of juxtaposition is tedious. There are various cases to consider, since cancellation of central inverses can proceed differently depending upon how words are grouped. Rather than enumerating the various cases and laboriously producing a separate proof of associativity for each, we will prove that $(W, *)$ and f form a free group on A using the so-called *van der Waerden trick*.

The basic idea is to consider the members of W , not as reduced words, but as permutations of reduced words. To facilitate this, we construct a kind of "scale model" of A , using permutations instead of alphabet letters. We use this scale model of A to construct a scale model of $(W, *)$, using compositions of permutations instead of words. And then we construct a scale model of f that relates the respective scale models of A and W in the same way that f relates A and W . Finally, we prove that the scale models of $(W, *)$ and f form a free group on the scale model of A ; and we infer from this that our "original" $(W, *)$ and f form a free group on our original A .

We begin with the general case and then illustrate with a simple example. To each letter $a \in A$ and each $\epsilon = \pm 1$ there corresponds a *single-letter prefixing function* $[a^\epsilon] : W \rightarrow W$, defined by $[a^\epsilon](w) = a^\epsilon * w$ for all $w \in W$. The function $[a^\epsilon]$ takes any reduced word $w \in W$ and juxtaposes it with a^ϵ . (Since juxtaposition is a binary operation on W , the word $[a^\epsilon](w)$ is guaranteed to be reduced, and so $[a]$ is a well-defined function from W to itself.) We then prove that $[a^\epsilon] \circ [a^{-\epsilon}] = 1_W = [a^{-\epsilon}] \circ [a^\epsilon]$ for each $a \in A$; and so it follows that each $[a^\epsilon]$ is a permutation on W with inverse $[a^{-\epsilon}]$. (Thus $[a^\epsilon]$ and $[a^{-\epsilon}]$ "undo" each other, and we have $[a^\epsilon]^{-1} = [a^{-\epsilon}]$.) We define $[A]$ as the set of all single-letter prefixing functions $[a]$ where $a \in A$. (This is our "scale model" of A .) Now, we know that the set of *all* permutations on $[W]$ is a group under composition of

functions; and we define $[W]$ as the subgroup of this group generated by $[A]$.¹⁹ (This set $[W]$ under composition of functions is our “scale model” of W under juxtaposition.) Thus $[W]$ is the set of permutations of W of the form $[a_1^{\epsilon_1}] \circ \dots \circ [a_n^{\epsilon_n}]$ where $a_i \in A$ and $\epsilon_i = \pm 1$. The members of $[W]$ are the permutations of reduced words obtainable by successive prefixing of alphabet letters and/or inverses of alphabet letters. We can think of these as prefixing functions more generally (including both single-letter and multi-letter prefixing functions).

Example: Where $A = \{a, b\}$, our single-letter prefixing functions are $[a], [b], [a^{-1}]$, and $[b^{-1}]$. (Note that $[a]$ and $[a^{-1}]$ are inverse functions, as are $[b]$ and $[b^{-1}]$.) Each of these can be applied to any reduced word: for example we have $[a](bba) = abba$ and $[b^{-1}](bba) = ba$. We thus have $[A] = \{[a], [b]\}$. And so $[W]$ contains all the prefixing functions generated by $[A]$, that is, all possible compositions of $[a], [b], [a^{-1}]$, and $[b^{-1}]$ (with repetitions allowed). For example we have $[a] \circ [a] \circ [b^{-1}] \in [A]$, which amounts to successive juxtaposition with b^{-1} , a and a , so that we have for instance $([a] \circ [a] \circ [b^{-1}])(ba^{-1}b) = ab$.

Note that in $[W]$ we do *not* have unique factorization into single-letter prefixings: different products of single-letter prefixings can yield the same overall function. (Example: $[a] \circ [a^{-1}] \circ [b] = [b] \circ [b^{-1}] \circ [b]$.) However, if we require that the product resulting from factorization corresponds to a *reduced* word, we do get uniqueness: for each $\sigma \in [W]$ there is a unique reduced word $a_1^{\epsilon_1} \dots a_n^{\epsilon_n}$ such that $\sigma = [a_1^{\epsilon_1}] \circ \dots \circ [a_n^{\epsilon_n}]$. We call this factorization the *reduced form* of σ . (Example: $[b]$ is the reduced form of $[a] \circ [a^{-1}] \circ [b]$.) The uniqueness of reduced forms will be important later.

We then define $[f]: [A] \rightarrow [W]$ such that $[f]([a]) = [a]$ for all $[a] \in [A]$. (This is our “scale model” of f .) Thus $[f]$ simply maps each single-letter prefixing function $[a] \in A$ to itself, considered as a prefixing function in $[W]$.

Next we show that $[W]$ and $[f]$ form a free group on $[A]$. (From this result regarding the “scale models” we’ll easily infer that W and f form a free group on A .) We see immediately that $[W]$ is a group under composition of functions: it’s a subgroup of the group of all permutations on W . (In particular, associativity is obvious, and we circumvent the tedious proof mentioned above.)

It remains to prove that for every group $(G, \cdot)$ and every function $g: [A] \rightarrow G$, there is a unique homomorphism $\phi: [W] \rightarrow G$ such that $g = \phi \circ [f]$. We proceed as follows. Let $(G, \cdot)$ be a group and $g: A \rightarrow G$ be a function. Define $\phi: [W] \rightarrow G$ such that for each $\sigma \in [W]$ we have:

$$\phi(\sigma) = g([a_1])^{\epsilon_1} \cdot g([a_2])^{\epsilon_2} \cdot \dots \cdot g([a_n])^{\epsilon_n}$$

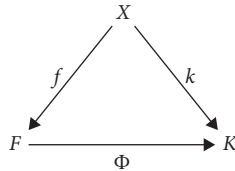
¹⁹ A set of generators $\{g_1, \dots, g_n\}$ is a set of group elements such that possibly repeated application of the generators on themselves and each other is capable of producing all the elements in the group. The set of generators is said to generate the relevant group.

where $[a_1^{\epsilon_1}] \circ [a_2^{\epsilon_2}] \circ \dots \circ [a_n^{\epsilon_n}]$ is the reduced form of σ . To apply ϕ to $\sigma \in [W]$, we first factorize σ into reduced form, then apply g to each factor individually, finally multiplying the results together in G . (The uniqueness of the reduced form ensures that ϕ is a well-defined function on $[W]$.) It follows easily enough that $g = \phi \circ [f]$, since if $[a] \in [A]$, then by the definition of $[f]$ we have $[f]([a]) = [a]$, and so $(\phi \circ [f])([a]) = \phi([a]) = g([a])$, by the definition of ϕ . Showing that ϕ is a homomorphism is more involved. For $\sigma_1, \sigma_2 \in [W]$, with corresponding reduced words $w_1, w_2 \in W$, there are two cases: either the concatenated word $w_1 w_2$ is reduced, or it isn't. If it is, then it follows quickly that $\phi(\sigma_1 \circ \sigma_2) = \phi(\sigma_1) \cdot \phi(\sigma_2)$. If not, then we use the Sandwich Lemma to write $w_1 w_2 = w_1' u u^{-1} w_2'$ where $w_1' w_2'$ is reduced. We can therefore apply the same reasoning as in the reduced case to show that $\phi(\sigma_1 \circ \sigma_2) = \phi(\sigma_1) \cdot \phi(\sigma_2)$. We then prove uniqueness: since any homomorphism Ψ such that $(\Psi \circ [f])([a])$ must agree with ϕ on the generating set $[A]$, it must also agree with ϕ on the whole of $[W]$. We therefore show that $[W]$ and $[f]$ form a free group on $[A]$.

Finally, exploiting the structural similarity between our “scale models” and our “originals”, we infer that W and f form a free group on A . Because each prefixing function has a unique reduced product, there's a bijective correspondence between prefixing functions and reduced words; and so the relationship between A , W , and f and mirrors that between $[A]$, $[W]$, and $[f]$. We thus see that W and f form a free group on A , as required. ■

5. Free Group Theorem: The “Abstract” Proof

Here we provide a different, more abstract, proof of the existence of free groups.²⁰ Recall the definition of a *Free Group*: Let X be a set. Group F is free on X if there is a map $f : X \rightarrow F$ and for any group K and map $k : X \rightarrow K$ there is a unique group homomorphism $\Phi : F \rightarrow K$ such that $k = \Phi \circ f$, that is, so that the following diagram commutes.



To prove the existence of free group F we will define two other groups on X , G_B and G_α , with respective maps g_B and g_α . For now, think of G_B as (roughly) the group composed of all groups on X (B for ‘Big’), and think of G_α as one G_B ’s components.

²⁰ This proof is due to Michael Barr (1972).

Inclusion Lemma (INC): Let G be a group with map $g : X \rightarrow G$. Then there is a subgroup H of G and map $h : X \rightarrow H$ such that h generates H and $g = j \circ h$ where j is the inclusion map.

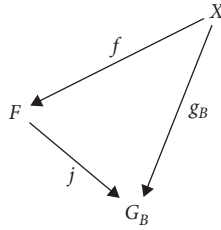
Definition of *generates*:

$h : X \rightarrow H$ generates $H \equiv$ the image $h(X)$ generates H

Set A generates group $H \equiv$ no proper subgroup of H contains A

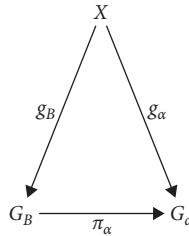
Proof of INC: Let H be the intersection of all subgroups of G that contain $g(X)$. No proper subgroup of H contains $g(X)$ so $g(X)$ generates H .

Given INC we say: H is the subgroup of G generated by $g : X \rightarrow G$. We define F as the subgroup of G_B that is generated by g_B . So the first triangle commutes.



The inclusion map $j : F \rightarrow G_B$ is a homomorphism since $j(x) \cdot j(y) = x \cdot y$ defines inclusion maps.

We now move to the middle triangle. We need no theorem to introduce the projection map. Its existence falls out of our definitions of G_B and G_α . Consider a collection of groups on X that are each generated by their corresponding maps. We then have a collection of pairs (G_α, g_α) where each map $g_\alpha : X \rightarrow G_\alpha$ generates its corresponding G_α . We now define G_B as the Cartesian product of such groups.



Definition: $G_B = \prod G_\alpha$

Example: Let $G_1 = \{a_1, b_1\}$ and $G_2 = \{a_2, b_2\}$, then

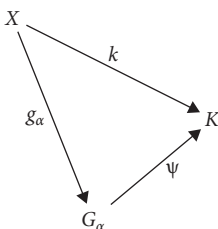
$G_1 \times G_2 = \{(a_1, a_2), (a_1, b_2), (b_1, a_2), (b_1, b_2)\}$ G_B is a group whose operation entails: $(a_1, a_2) \cdot (b_1, b_2) = (a_1 b_1, a_2 b_2)$, where $a_i b_i$ is the product in G_i .

Definition: $g_B = \prod g_\alpha$

Example: Let $g_1 : X \rightarrow G_1$ and $g_2 : X \rightarrow G_2$ be maps, then $g_1 \times g_2 : X \rightarrow (G_1 \times G_2)$, such that: $(g_1 \times g_2)(x) = (g_1(x), g_2(x))$. So $g_B : X \rightarrow G_B$ is a map.

Now we define our projection map. Let $\Pi_\alpha : G_B \rightarrow G_\alpha$ be a projection map. Example: $\Pi_1 : G_B \rightarrow G_1$ (obviously a homomorphism). So it is clear that: $g_\alpha = \Pi_\alpha \circ g_B$ and the middle triangle commutes.

We now consider the third triangle. Recall: given set X there is a collection of pairs (G_α, g_α) where each G_α is a group and $g_\alpha : X \rightarrow G_\alpha$ generates G_α .



Isomorphism Lemma (ISO): if K is any group and $k : X \rightarrow K$ generates K , then for some α there is an isomorphism Ψ on G_α onto K such that $k = \Psi \circ g_\alpha$.

Intuitive Proof of ISO: Simply take all pairs (G_α, g_α) where G_α is a group and g_α generates G_α . Then K is included by definition. That's the intuitive idea but the problem is that "all" leads to "the usual logical paradoxes". So we turn to a more rigorous poof but first we need some lemmas.

Sub-lemmas for Rigorous Proof of ISO:

Sub-lemma SL1: Let $k : X \rightarrow K$ generate K . Then $|K| \leq \max(|X|, \aleph_0)$ (proved on pp. 366–7 of Barr 1972).

Sub-lemma SL2: Y is a set. Then the collection of groups G generated by Y has cardinality less than or equal to $|Y|^{\aleph_0}$. (Just consider the possible ways of filling out the relevant multiplication table.)

Rigorous proof of ISO: For each cardinal s with $s \leq \max(|X|, \aleph_0)$ choose a set Y_s with $|Y_s| = s$. Consider the set of all groups G with underlying set Y_s ; by SL2 Y_s exists. Consider all pairs (G_α, g_α) where $g_\alpha : X \rightarrow G_\alpha$ generates G_α . ISO follows from this construction and SL1.

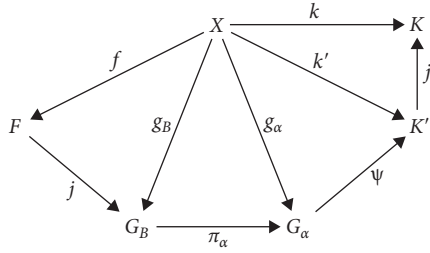
We now have our three commuting triangles. We put them together to see that the large diagram commutes. But we must consider two separate cases, the second of which requires a special qualification.

Recall the definition of F : Subgroup of G_B generated by g_B . By INC: $g_B = j \circ f$.

Case 1 (already depicted in the previous large diagram): Assume $k : X \rightarrow K$ generates K . By ISO there is an α and an isomorphism Ψ such that $k = \Psi \circ g_\alpha$ (third triangle). Note: $g_\alpha = \pi_\alpha \circ j \circ f$ (combining first and second triangles). Then

by substitution: $k = \Psi \circ \pi_\alpha \circ j \circ f$ (combining all three). But the composite $\Psi \circ \pi_\alpha \circ j$ is a homomorphism in virtue of its components being homomorphisms. So let $\Phi = \Psi \circ \pi_\alpha \circ j$. Case 1 is thus closed.

Case 2 (depicted below): Assume $k : X \rightarrow K$ does not generate K . By INC there is a subgroup K' of K such that $k' : X \rightarrow K'$ generates K' and $k = j' \circ k'$, where $j' : K' \rightarrow K$ is an inclusion map. We know there is a homomorphism $\Phi' : F \rightarrow K'$, such that $k' = \Phi' \circ f$ (see case 1). So let $\Phi = j' \circ \Phi'$ since $j' : K' \rightarrow K$ is a homomorphism. Then $\Phi : F \rightarrow K$ is a homomorphism and $k = \Phi \circ f$. Case 2 is also closed.



Uniqueness of Φ

Let: $\Phi_1 : F \rightarrow K$ and $\Phi_2 : F \rightarrow K$ be homomorphisms, where $\Phi_1 \circ f = \Phi_2 \circ f = k$. Let F_0 be the set of x in F such that $\Phi_1(x) = \Phi_2(x)$. But then F_0 is a subgroup of F . Now we prove that $f(X) \subseteq F_0$. For all $x \in X : \Phi_1 \circ f(x) = \Phi_2 \circ f(x)$ (by def. of Φ_i). $\Phi_1(f(x)) = \Phi_2(f(x))$ (by def. of $\circ$). But $f(X)$ generates F (because F is free). So $F_0 = F$, hence $\Phi_1 = \Phi_2$. The proof is complete. ■

6. The Explanatory Value of the Proofs

6.1 The constructive proof and local dependence-based explanatory value

The constructive proof explains the existence of free groups by building them and showing how the intrinsic structure of what's built guarantees the universal property. For example, the proof builds “words” and “letters” from members of an arbitrary set, before defining a suitable group operation. The proof then shows how the universal property depends on features of this construction. This structural feature of the proof appears to fit the dependence-based model of explanation in the philosophy of science. In that case, assuming that a proof has explanatory value if it fits this model of explanation, then the constructive proof has a distinctive kind of explanatory value.

The dependence-based model of explanation really breaks into two distinct but analogous models. On the one hand there are dynamic causal theories on which (roughly) a phenomenon is explained by describing the cause that the phenomenon causally depends on.²¹ On the other hand there are synchronic reductive theories on

²¹ For example, see Lewis (1986).

which (roughly) a phenomenon is explained by describing the underlying structure or process that the phenomenon reduces to, or metaphysically depends on (Dizadji-Bahmani et al. 2010). The reductive theories are somewhat closer to what we are after. Such theories try to model what is happening, for example, in the reduction of the laws of thermodynamics to statistical mechanics, and in the reduction of rigid body mechanics to particle mechanics.

In the former case, hypothetical statistical mechanical systems are constructed, and we are shown how the principles of thermodynamics fall out of these constructions plus the statistical mechanical laws. To explain the Boyle-Charles Law, for example, one constructs an idealized gas and describes it in terms of Newton's laws. One then shows (by deduction) that the mean kinetic energy of the gas particles gives rise to the Boyle-Charles Law, since it can be deduced by identifying temperature with mean kinetic energy (Dizadji-Bahmani et al. 2010: §2). In the latter case, one constructs a hypothetical idealized microphysical system, and shows how the principles of rigid-body mechanics fall out of these constructions plus the microphysical laws. To explain the principle of mass additivity, for example, one constructs an idealized microphysical system and describes it in terms of Newton's laws. One then shows (by deduction) that the mass of its composite is the sum of the masses of its elementary components (McQueen 2015).

The constructive proof of free groups does something very similar, and so should therefore be similarly described by reductive models of explanation. For in this proof, a group is constructed out of a set, and we are shown how the universal property definitive of free groups falls out of this construction plus principles of group theory.²² We believe that this is at least similar to Steiner's (1978a) idea that (i) to explain the behaviour of an entity, one deduces the behaviour from the essence or nature of the entity and (ii) mathematical proofs exhibit this deductive structure.²³ Thus, already existing models of explanation give us good reason to think the constructive proof is explanatory, since the constructive proof satisfies the key requirements of such models.²⁴

The abstract proof does not do this. Although we have a construction given in terms of the Cartesian product of all groups on a set, we are given no information about the intrinsic structure of these component groups. Instead the abstract proof works with abstract relationships among groups to show how those relationships guarantee that (the subgroup of) this Cartesian product satisfies the universal property. But we are left

²² Perhaps talk of "falls out" sounds a bit loose and does not get at the core distinction between a proof and an explanatory proof. But here we don't mean simply that the universal property logically follows from the construction in question, rather, we mean that the universal property naturally arises from the core properties of the construction in question.

²³ This is also similar to Colyvan's (2012: ch. 5) suggestion that relevance (in the technical sense) might be a way of spelling out this local/intrinsic notion of explanation in mathematics.

²⁴ Here we're arguing by analogy. We're appealing to accepted similarities between two things (a reductive explanation and a constructive proof) to support the conclusion that some further similarity between them exists (namely, explanatory value). Of course Steiner's account of mathematical explanation has its critics (e.g., see Mancosu 2008a, 2008b) but to be clear, we are not suggesting that Steiner's account is correct or problem free. We are merely noting that there is an explanatory virtue found in the constructive proof that might be fruitfully spelled out along similar lines to at least some parts of Steiner's account.

wondering what it is about the intrinsic structure that guarantees this.²⁵ And so if a proof can only have explanatory value if it is modelled by a dependence account, then the abstract proof may be seen to be unexplanatory. But this is too quick. There are explanatory virtues in the abstract proof, but they are apparently of a different kind.

6.2 *The abstract proof and global unification-based explanatory value*

Consider a different kind of explanatory virtue based on the unificationist account of explanation found in the philosophy of science (Friedman 1974; Kitcher 1981). On the unification approach in the philosophy of science, an event is explained by deriving the occurrence of the event using a theory that unifies many diverse phenomena, and thereby showing that the event is part of a very general, perhaps utterly pervasive, pattern of events in the universe. In the best-developed unification account of explanation, due to Philip Kitcher, an event is explained by deducing it using the theory that unifies the phenomena better than any other.²⁶

One can straightforwardly adapt this philosophy of science account to the philosophy of mathematics by replacing “event” and “occurrence of event” with “theorem”, meanwhile “theory” can be replaced by “proof”. For example, on the unificationist approach in the philosophy of mathematics, a theorem is explained by deriving the theorem using a proof that unifies many diverse theorems, and thereby showing that the theorem is part of a very general, perhaps utterly pervasive, pattern of theorems in mathematics. With this in mind, consider what Michael Barr says about the abstract proof:

The proof is modelled after that of the general adjoint functor theorem of category theory and, as such, is readily adapted to solving any universal mapping problem in the category of groups, such as the existence of free products. It also works in any category consisting of all the algebras and algebra homomorphisms of any algebraic theory. [...] Thus included are all such categories as sets, sets with a base point (and base-point preserving functions), groups, abelian groups, rings, commutative rings, Lie rings, Jordan rings, algebras of these types, etc., each considered as a category with the evident definition of homomorphism. (Barr 1972: 364)

We can therefore say that the free group theorem is explained by the abstract proof because the abstract proof unifies many diverse free object existence theorems, and thereby shows that the free group theorem is part of a very general, persuasive, pattern of theorems in mathematics (free object theorems).

We need not think of the unification theory as being *the* theory of explanation, just as we need not think of the dependence theory as being *the* theory of explanation. We only need to think of them as providing a means of spelling out a source of explanatory value. These sources need be neither necessary nor sufficient conditions for possessing explanatory power. If that's right, then since the abstract proof fits so nicely into the

²⁵ This line of thought was expressed by some mathematicians and physicists in our informal discussions on the Physics Forum.

²⁶ This particular formulation is due to Michael Strevens (2004).

philosophy of mathematics version of the unification account, then arguably the abstract proof has distinctive explanatory value.²⁷

Perhaps the abstract proof also has some claim to exhibiting the reduction-style virtues. This may well be but if so, such virtues are not prominent. The abstract proof does not build a specific group and prove that there must be a free group from those specifics. That's why it's not meant to be analogous to reductive explanations, which (e.g., in the rigid-body mechanics case) build a specific microphysical system and deduce a given property (mass additivity, say) from those specifics. Rather, the relevant property (freeness) is derived from certain very general properties of groups and this becomes explanatory insofar as the proof can enable one to see how it can be generalized to other domains, thereby unifying them. But it's plausible that these things come in degrees. Perhaps the abstract proof is somewhat reductive, even if that's not the feature of it that yields most of its explanatory value, and perhaps the constructive proof is somewhat unificatory even if that's not the feature of it that yields most of its explanatory value.

6.3 Which proof is more explanatory?

If the above is a correct characterization of the situation, then the two proofs are hard to compare because their primary sources of explanatory value differ. For example, if both proofs were modelled only by the unificationist model, we could ask which proof has the more general scope, or which proves a range of theorems using the most minimal basic assumptions (or both). Determining which is more explanatory wouldn't exactly be straightforward but we would at least have a clear way forward. But it is not clear how to compare the relative explanatory strengths of proofs whose primary explanatory value come from such distinct sources.

What do mathematicians think? Some comments from Saunders Mac Lane on this issue are interesting. Mac Lane notes that one of the applications of the category theory Representability Theorem is that it facilitates a neat (category theory) proof of the Free Group Theorem "without entering into the usual (rather fussy) explicit construction of the elements of [the free group on X] as equivalence classes of words in letters of X" (Mac Lane 1998: 123). It's not clear from Mac Lane's comments whether this claimed advantage for the abstract approach to proving the Free Group Theorem is supposed to be an *explanatory* advantage. At first blush the advantage looks pragmatic: the abstract proof avoids some tedious constructions and saves some ink. But the claimed advantage can also be thought of as an explanatory advantage: whereas the constructive proof bogs down in detail, the category theory proof rises above such details to reveal the real reasons for the existence of free groups. According to this line of thought, the real reason for the existence of free groups is found at the more abstract structural

²⁷ In particular, given that we are not committed to the unification account being *the* account of explanation, some of the noted shortcomings of unificationism need not concern us. Indeed, the problem cases for Kitcher's account do not undermine it completely but, rather, serve to highlight its limitations as the complete account of explanation. (See Tappenden 2005 for discussion of this issue.)

level. The existence of free groups is just a special case of a more general result, so focusing on the details in group theory is to miss the point, or so the suggestion goes.²⁸ A related suggestion is that a proof of free groups is explanatory to the extent that it helps justify/secure/illuminate the applications of free groups in group theory. If this is on the right track, it may provide a neutral way of comparing the explanatory power of proofs whose respective explanatory values come from different sources (e.g., local dependence versus global unification).

So we can ask, to what extent is the subgroup of G_B generated by g_B that which one has in mind when engaged in applications? Analogously: to what extent is the free group with group operation juxtaposition that which one has in mind when engaged in applications? Does it ever make a difference? These are not questions we can answer here but these are the kinds of questions that need to be addressed in advancing our understanding of mathematical explanation.²⁹

Our conclusion may seem a little unsatisfying: there is good reason to suspect that there are two competing candidates for explanatory power in mathematics—two flavours of mathematical explanation, if you like—and it is difficult to make trade-offs between the two. But as we said at the outset, mathematical explanation is puzzling—puzzling enough that we should be suspicious of any account that promises easy answers. In any case, we make no apologies for not offering easy answers. Instead, we offer a case study that we believe is helpful in shedding light on the nature(s) of mathematical explanation. We have argued that a given theorem admits two intuitively explanatory proofs, one which is structurally similar to reductive explanation, another which is structurally similar to unificationist explanations. We speculate that the explanatoriness derives from these structures.

Although it is common to talk of a proof being explanatory or not, and we too mostly follow this way of talking, it seems to us that it is more plausible that explanatory virtues come in degrees. Those proofs that exhibit an explanatory virtue to a high degree are those that we speak of as being explanatory. (Just as belief comes in degrees and if the degree is high enough we tend to treat that as full belief.) But accepting that explanatory virtue comes in degrees and that there is more than one kind of explanatory virtue does not trivialize the view. It does not, for example, mean that all proofs are explanatory because they are all explanatory to some degree in some explanatory virtue or other. The proofs in question need to exhibit the explanatory virtue(s) to a

²⁸ This line of thought was also expressed by some of the mathematicians and physicists on the Physics Forum discussion.

²⁹ In this spirit, here are a couple of specific applications to think about:

- (i) Often one proves a result about groups by first establishing the result for free groups and then showing how it holds for the quotient of these groups. When these groups are abelianized (mod out by commutators) this has important consequences for computing things like Ext and Tor.
- (ii) Every group is a quotient of two free groups. (Let G be any group and let F_G be the free group generated by the elements of G . The universal property of this free group provides a homomorphism $F_G \rightarrow G$ and let K denote its kernel. By the first isomorphism theorem it follows that $F_G / K = G$ and since subgroups of free groups are free, this establishes that every group is a quotient of two free groups.) This entails that every group has a presentation. (The generators are given by the generators of F_G and the relations are given by the generators of K .)

suitably high degree. What is a high enough degree? This is probably context sensitive and perhaps also vague, but there will be clear cases on either side. There will also be some difficult comparisons—even among the clear cases of explanatory proofs. Indeed, the two proofs in this chapter illustrate such difficulties: one proof is high in the unificatory stakes and low in the reductionist stakes (the abstract proof), the other is high in the reductionist stakes and low in the unificatory stakes (the constructive proof). But according to our account, both proofs are explanatory, albeit explanatory for different reasons. Each is explanatory because it exhibits one of the explanatory virtues to a high degree but it is not clear how to compare these two kinds of explanatory virtues so there is no straightforward way to say which proof is the more explanatory. Indeed, there may be no fact of the matter about such comparisons.

As we have already noted, further philosophical work needs to be done on understanding the broader roles of free groups in mathematics to see which of the proofs of the theorems in question best support these roles. We also need to look at proofs of theorems from a variety of areas of mathematics to see if the same issues arise. Finally, we need greater collaboration between mathematicians and philosophers on this project. This is not something philosophers can do alone. Most philosophers' intuitions about explanatory power in mathematics run out fairly quickly and, in any case, are unlikely to be reliable. Our case study of the Free Group Theorem is just a small step towards a better understanding of the intricacies of the explanatory virtues of different proofs.

Acknowledgements

We are grateful to Sam Baron, Rachael Briggs, Clio Cresswell, Ed Mares, Daniel Nolan, Jeff Pelletier, Graham Priest, Dave Ripley, and Jamie Tappenden for discussions about the material covered in this chapter. We are also grateful to several mathematicians and physicists who contributed to our discussions on the Physics Forum. Their insights about explanation in mathematics were extremely helpful, as was their suggestion of looking at the Free Group Theorem. Material from this chapter was presented to the 2014 Australasian Association of Logic Conference at the University of Sydney. We are grateful to the audience in attendance for their very helpful comments and suggestions. We'd also like to acknowledge Manya Raman-Sundström and the anonymous referees for this volume for many helpful comments on earlier versions of this chapter. This work was funded by an Australian Research Council Future Fellowship grant to Mark Colyvan (grant number FT110100909).

References

- Aigner, M. and Ziegler, G. M. (2010), *Proofs from THE BOOK*, 4th edn. (Heidelberg: Springer).
 Baker, A. (2005), 'Are There Genuine Mathematical Explanations of Physical Phenomena?', *Mind* 114: 223–38.
 Baker, A. (2009), 'Mathematical Explanation in Science', *British Journal for the Philosophy of Science* 60: 611–33.

- Baker, A. (2010), 'Mathematical Induction and Explanation', *Analysis* 70: 681–9.
- Baron, S. (2014), 'Optimization and Mathematical Explanation: Doing the Lévy Walk', *Synthese* 191: 459–79.
- Baron, S. and Colyvan, M. (2016), 'Time Enough for Mathematical Explanation', *Journal of Philosophy* 113: 61–88.
- Baron, S., Colyvan, M., and Ripley, D. (2017), 'How Mathematics Can Make a Difference', *Philosophers' Imprint* 17:1–19.
- Barr, M. (1972), 'The Existence of Free Groups', *American Mathematical Monthly* 79: 364–7.
- Colyvan, M. (2001), *The Indispensability of Mathematics* (New York: Oxford University Press).
- Colyvan, M. (2002), 'Mathematics and Aesthetic Considerations in Science', *Mind* 111: 69–74.
- Colyvan, M. (2010), 'There is No Easy Road to Nominalism', *Mind* 119: 285–306.
- Colyvan, M. (2012), *An Introduction to the Philosophy of Mathematics* (Cambridge: Cambridge University Press).
- Colyvan, M. (forthcoming), 'The Ins and Outs of Mathematical Explanation', *Mathematical Intelligencer*.
- Davis, P. J. and Hersh, R. (1981), *The Mathematical Experience* (Boston: Birkhäuser).
- Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. (2010), 'Who's Afraid of Nagelian Reduction?', *Erkenntnis* 72: 303–22.
- Friedman, M. (1974), 'Explanation and Scientific Understanding', *Journal of Philosophy* 71: 5–19.
- Giaquinto, M. (2016), 'Mathematical Proofs: The Beautiful and the Explanatory', *Journal of Humanistic Mathematics* 6: 52–72.
- Gowers, T. and Neilson, M. (2009), 'Massively Collaborative Mathematics', *Nature* 461: 879–81.
- Hafner, J. and Mancosu, P. (2005), 'The Varieties of Mathematical Explanation', in P. Mancosu, K. F. Jørgensen, and S. A. Pedersen (eds.), *Visualization, Explanation and Reasoning Styles in Mathematics* (Dordrecht: Springer), 215–50.
- Hardy, G. H. (1967), *A Mathematician's Apology* (Cambridge: Cambridge University Press).
- Inglis, M. and Aberdeen, A. (2015), 'Beauty Is Not Simplicity: An Analysis of Mathematicians' Proof Appraisals', *Philosophia Mathematica* 23: 87–109.
- Kitcher, P. (1981), 'Explanatory Unification', *Philosophy of Science* 48: 507–31.
- Lange, M. (2009), 'Why Proofs by Mathematical Induction are Generally Not Explanatory', *Analysis* 69: 203–11.
- Lange, M. (2014), 'Aspects of Mathematical Explanation: Symmetry, Unity, and Salience', *Philosophical Review* 123: 485–531.
- Lange, M. (2016), 'Explanatory Proofs and Beautiful Proofs', *Journal of Humanistic Mathematics* 6: 8–51.
- Lewis, D. (1986), 'Causal Explanation', in *Philosophical Papers*, vol. II (New York: Oxford University Press), 214–40.
- Lyon, A. and Colyvan, M. (2008), 'The Explanatory Power of Phase Spaces', *Philosophia Mathematica* 16: 227–43.
- Mac Lane, S. (1998), *Categories for the Working Mathematician*, 2nd edn. (New York: Springer).
- McQueen, K. J. (2015), 'Mass Additivity and A Priori Entailment', *Synthese* 192: 1373–92.
- Mancosu, P. (2008a), 'Mathematical Explanation: Why it Matters', in P. Mancosu (ed.), *The Philosophy of Mathematical Practice* (Oxford: Oxford University Press), 134–49.

- Mancosu, P. (2008b), 'Explanation in Mathematics', *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.). <<http://plato.stanford.edu/archives/fall2008/entries/mathematics-explanation/>>.
- Pincock, C. (2015), 'The Unsolvability of the Quintic: A Case Study in Abstract Mathematical Explanation', *Philosophers' Imprint* 15: 1–19.
- Raman-Sundström, M. and Öhman, L. (forthcoming), 'Mathematical Fit: A Case Study', *Philosophia Mathematica*.
- Resnik, M. and Kushner, D. (1987), 'Explanation, Independence, and Realism in Mathematics', *British Journal for the Philosophy of Science* 38: 141–58.
- Rotman, J. J. (1965), *The Theory of Groups: An Introduction* (Boston: Allyn & Bacon).
- Steiner, M. (1978a), 'Mathematical Explanation', *Philosophical Studies* 34: 135–51.
- Steiner, M. (1978b), 'Mathematics, Explanation, and Scientific Knowledge', *Noûs* 12: 17–28.
- Strevens, M. (2004), 'The Causal and Unification Approaches to Explanation Unified—Causally', *Noûs* 38: 154–76.
- Tappenden, J. (2005), 'Proof Style and Understanding in Mathematics I: Visualization, Unification and Axiom Choice', in P. Mancosu, K. F. Jørgensen, and S. A. Pedersen (eds.), *Visualization, Explanation and Reasoning Styles in Mathematics* (Dordrecht: Springer), 147–214.

12

When are Structural Equation Models Apt? Causation versus Grounding

Lina Jansson

1. Introduction

The notion of ground has made a prominent rise in contemporary metaphysics. While much about the notion of ground remains under debate, one feature has reached near consensus.¹ Ground is closely connected to a form of explanation. As Jenkins (2013) points out, we often use explanatory locutions such as “because” and “in virtue of” when discussing ground. The association to explanation is often explicit as in Dasgupta’s discussion of ground:

Imagine you are at a conference, and imagine asking why a conference is occurring. A causal explanation might describe events during the preceding year that led up to the conference: someone thought that a meeting of minds would be valuable, sent invitations, etc. But a different explanation would say what goes on make the event count as a conference in the first place. Someone in search of this second explanation recognizes that conferences are not *sui generis*, so that there must be some underlying facts about event *in virtue of which* it counts as being a conference, rather than (say) a football match. Presumably it has something to do with how the participants are acting, for example that some are giving papers, others are commenting, and so on. An answer of this second kind is a statement of what *grounds* the fact that a conference is occurring. (Dasgupta 2014: 3)

Even if we accept that the notion of ground is tied to the notion of explanation, there are substantive open questions about how exactly we should understand the connection to explanation.²

Recently, Schaffer (2016) and Wilson (2016 and forthcoming) have argued that ground should be understood as an explanation-backing relation akin—or identical—to causation. Just as causal relationships can back a particular type of explanation, causal

¹ See Wilson (2014: 555–6) for a dissenting view.

² See Bliss and Trogdon (2014) for a delineation of several of these issues.

explanation, so grounding relationships can back a particular type of explanation, grounding explanation. Part of their argument is based on the claim that both notions can be productively treated using the tools of structural equations and directed graphs. The use of directed graphs and structural equations in an account of explanation is most closely associated with Woodward's (2003) interventionist counterfactual account of causal explanation. If we can use a similar framework in the grounding case, then we could unify the structure of grounding explanations and that of causal explanations, thereby adding weight to the idea that ground should be treated as an explanation-backing relation akin to causation. In section 2 I will briefly review how structural equation modelling works in the causal case before showing how the formal framework can be extended to the grounding case.

In section 3 I will argue that it is only the formal framework that carries over to the grounding case. In particular, the *seeming* unification of the structural equations approach to explanation disappears once we take into account what it takes for a structural equations model to have appropriately captured the situation that we are modelling. As Schaffer (2016) and Blanchard and Schaffer (2017) emphasize, structural equation modelling is a type of *modelling*. Once we are given a model of some scenario or system, the obvious question to ask is whether the model is any good. That is, is the model an apt or fitting one (for the purpose at hand)? For a model to be a good one for the purposes at hand it has to contain appropriate (whatever that turns out to mean) variables and appropriately (whatever that turns out to mean) represent the relations of causal or grounding relevance.

I will show that there are important differences between the causal case and the grounding case in how we can judge whether or not a model is an appropriate representation of a scenario. Once we take into consideration the aptness conditions associated with causal explanations we find that we rely on non-trivial a posteriori knowledge about the nature of the specific causal processes relevant to the model. This allows us to rule out some models as failing to adequately represent the causal dependencies of interest. In the grounding case we do not have access to such specific a posteriori knowledge of the grounding processes relevant to the grounding models. This means that we cannot draw on the same resources to rule out certain models as failing to appropriately capture the grounding dependencies of interest.

Before I set out my argument, let me address a natural initial objection. The question of whether a model is an apt one is a question with epistemic and pragmatic overtones. That such concerns should be part of the modelling criteria for ground and grounding explanation at all might seem surprising given that ground is supposed to be a metaphysical notion. It is tempting to try to avoid epistemic concerns by simply stipulating that an apt model is one that correctly captures the grounding structure.³ However, there is a high cost to going down this route.

³ I take Wilson (2016) to choose this option. Schaffer (2016: 68–9) discusses the worry of the lack of a uniquely appropriate model but does not offer a solution.

The connection to explanation is often viewed as the strongest reason to accept the existence of a grounding relation that is conceived as some kind of determination relation along the lines of causation: whether this determination relation is a productive relation, a generative relation, or a dependence relation.⁴ Audi (2012: 105) formulates this line of argument particularly clearly.

- (1) If one fact explains another, then the one plays some role in determining the other.
- (2) There are explanations in which the explaining fact plays no causal role with respect to the explained fact.
- (3) Therefore, there is a non-causal relation of determination.

This argument runs from the existence of what are supposed to be clear cases of metaphysical, non-causal, explanation to the postulation of a relation of ground. In order for such an argument to work, we have to recognize the explanations in question as explanations. So, whatever the explanation-backing relation (if any) is, it should be a relation that we can recognize as obtaining. If it is not a relation that we have epistemic access to—at least in the sense of having good reason to think that the relation obtains—then we must either take ourselves to not have good reason to think that we have an explanation in the cases under consideration, or take it that the relation obtaining is not crucial for having those explanations after all. Either option would undermine the argument from the recognition of these explanations to the postulation of a relation of ground that is backing the explanations. Given these background assumptions, if we expect ground to be a relation that we can capture using structural equation models, then we should also expect to be able to use these models to recognize certain specific explanations. The models cannot simply be apt, but must be such that we can recognize that they are apt.⁵ This means that epistemic concerns about how we assess the aptness of grounding models cannot simply be set aside.

2. The Formalism of Structural Equations Models

In this section I will introduce the formal machinery associated with structural equations models. I will first present the causal case before showing how Schaffer and

⁴ See, for example, Schaffer (2016: 58) and, an early example, Kim (1974). Kovacs (2017) calls the argument from explanation to ground the “master argument” for grounding.

⁵ This already raises a worry about the argumentative strategy at hand. Insofar as we are concerned with grounding as a pure metaphysical relation—unadulterated by epistemic and pragmatic constraints—we have reasons to be concerned about whether structural equation models and the associated account of explanation are suited to provide this. We should be unsurprised to find that that relations modelled with an eye to explanations available to us are a mixture of ontic and epistemic considerations. (See Illari 2013 for an argument for this in relation to mechanistic explanations.) Woodward’s (2003: 179) interventionist counterfactual approach to explanation incorporates strong epistemic constraints. These constraints are reflected in a relative disinterest in the fundamental ontology. Importantly, the causal relations are not even taken to be good candidates for fundamental relations. A variable is taken to be a direct (or contributing) cause for another variable only relative to a variable set.

Wilson modify the notions associated with the causal case in order to apply a structural equations framework to the grounding case.

2.1 *Structural equations modelling of causal scenarios*

In Woodward's (2003) framework, causal relationships are represented using directed graphs, where a directed edge from one variable to another represents a direct causal relationship. In addition to the directed graph we also specify a collection of structural equations that specify exactly how the values of the dependent variables are functions of their direct causes.

I make use of two devices to represent causal relationships. A *directed graph* is an ordered pair $\langle \mathbf{V}, \mathbf{E} \rangle$ where $\mathbf{V}$ is a set of vertices that serve as the variables representing the relata of the causal relation and $\mathbf{E}$ a set of directed edges connecting these vertices. A directed edge from vertex or variable X to vertex or variable Y means that X *directly causes* Y ... The basic idea is that X is a direct cause of Y if and only if the influence of X on Y is not mediated by any other variables in the system of interest $\mathbf{V}$ in the following sense: there is a possible manipulation of X that would change the value of Y ... when all other variables in $\mathbf{V}$ are held fixed at some set of values in a way that is independent of the change in X ...

It will also be useful to represent causal relationships by means of systems of equations... in which each endogenous variable Y (i.e., each variable that represents an effect) may be written as a function of all and only those variables that are its direct causes... (Woodward 2003: 42)

Let me use a very simple example to illustrate the framework. Let us say that I want to model a scenario where a boulder falls and starts rolling towards a hiker. The hiker sees the falling boulder and ducks.⁶ To model the scenario we introduce a variable for whether or not the boulder falls, let us call it F , and a variable for whether or not the hiker ducks, let us call it D . For each variable we stipulate that it takes one of two values; 1 if the boulder falls and if the hiker ducks (respectively), and 0 otherwise. In order to represent the relations of causal relevance, we can make use of directed edges. The fall (or non-fall) is directly causally relevant to the ducking or non-ducking of the hiker. A causal diagram captures this simple causal structure.



We also need to represent the endogenous variables—in this case only D —as a function of their direct causes. This is easy to do. The hiker ducks if the boulder falls, but not otherwise. So, D takes the value 1 when F takes the value 1 and the value 0 when F takes the value 0; $D = F$.

Woodward's account is a non-reductive one. In the full account, the notion of a 'possible manipulation of X that would change the value of Y ... when all other variables in $\mathbf{V}$ are held fixed at some set of values in a way that is independent of the change in X ...'

⁶ We will expand on this case in section 3.

relies on a technical notion of an intervention on X with respect to Y . The notion of an intervention is itself defined in causal terms relative to a causal graph.⁷

2.2 *Structural equations modelling of grounding scenarios*

The cases discussed as examples of grounding explanations are of heterogeneous nature but often include examples like the following:⁸

- The fact that P and the fact that Q jointly ground the fact that $P \wedge Q$.
- The fact that $P \wedge Q$ does not ground the fact that P .
- The fact that Socrates exists grounds the fact that the singleton set containing Socrates as its sole member exists.
- The fact that the singleton set containing Socrates as its sole member exists does not ground the fact that Socrates exists.
- The fact that an act is virtuous grounds the fact that the gods love it.
- The fact that the gods love an act does not ground the fact that the act is virtuous.

All of the above seem like cases where there is a *directed* notion of dependence in place. Just as in the causal literature, it seems hard to analyse this directed notion of dependence purely in counterfactual terms. Some counterfactuals seem relevant, but others do not.⁹ As Wilson (2016) points out, we need to rule out the counterfactuals that seem to track back along the direction of causation and back along the direction of ground. It seems true that had Socrates not existed, then the singleton containing Socrates would also not have existed: this seems like the right counterfactual for the grounding case. It also seems true that had the singleton containing Socrates not existed, then Socrates would not have existed; this seems like the wrong counterfactual for the grounding case. The hope is that the grounding case could—like the causal case—make use of a non-reductive but informative account to clarify which counterfactuals are relevant.

In particular, notice that in the causal case we start by selecting an appropriate causal model where we specify the endogenous and exogenous variables. The formal notion of intervention is defined relative to such a model. It is, therefore, tempting to take the same strategy in trying to capture the relevant counterfactuals in the grounding scenario, and to capture the interventionist counterfactuals *in terms of* directed graphs and structural equations. This is the strategy of Schaffer (2016) and Wilson (2016). Wilson (2016: 3) characterizes an intervention by reference to a directed graph as “... a ‘clean’

⁷ See Woodward (2003: 98).

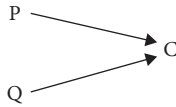
⁸ I should note that there is no consensus on what the proper relata of ground are. Here I am following the convention making the relata facts, but nothing that I will go on to say hinges on this choice. I will drop the reference to facts when it makes sentences less cumbersome to do so.

⁹ In all of the cases above, the antecedent would be a counterfactual and not a counterpossible. However, counterpossible antecedents will be needed for some examples in the grounding literature. For example, it may be taken to be the case that the fact that 2 exists grounds the fact that $\{2\}$ exists, but not vice versa.

alteration of the value of a particular variable that does not affect the values of upstream causal variables”.

To see the framework in action, let us return to the case of the fact that P together with the fact that Q grounding the fact that $P \wedge Q$.¹⁰ We introduce a variable for whether or not P obtains; let us call it P . Similarly, we introduce a variable for whether or not Q obtains; let us call it Q . Finally we introduce a variable for whether or not $P \wedge Q$ obtains; let us call it C . As in the previous case, each variable takes one of $\{0, 1\}$ depending on whether or not the fact in question obtains.

In order to represent the relations of grounding relevance, we can make use of directed edges. Whether or not P obtains is directly grounding relevant to whether or not $P \wedge Q$ obtains. Similarly, whether or not Q obtains is directly grounding relevant to whether or not $P \wedge Q$ obtains. Just as in the causal case, a directed graph does not tell us how the endogenous variables depend on their direct grounds (or causes). To specify this we make use of structural equations. We only need one structural equation in order to specify how $P \wedge Q$ depends on P and Q ; $C = \text{Min}(P, Q)$. That is, the value of C is equal to the lowest value taken by either of P or Q (or both).



According to the grounding model, some interventions on P , with Q fixed, entail a “downstream” alteration of C . In particular, if Q takes the value 1, an intervention that changes the value of P from 1 to 0 will have the downstream consequence of changing C from 1 to 0. However, P is not downstream from C , so no conclusions about changes in P based on interventions on C are supported by the grounding model.

So far this is just a formal characterization. We can still ask what the intervention in question represents. This brings with it the worry that interventions on variables in the grounding case are not “clean”. In particular, the variables involved (such as whether $P \wedge Q$ holds and whether P holds) are not *independently fixable* (IF). That is, they will clearly violate the criterion specified in Woodward (2015).

(IF): a set of variables V satisfies independent fixability of values if and only if for each value it is possible for a variable to take individually, it is possible (that is, “possible” in terms of their assumed definitional, logical, mathematical, mereological or supervenience relations) to set the variable to that value via an intervention, concurrently with each of the other variables in V also being set to any of its individually possible values by independent interventions.

(Woodward 2015: 316)

Woodward does not endorse IF as a general requirement on interventionist accounts. Rather, he takes it to indicate a difference between causal and non-causal dependence

¹⁰ This case is adapted, with some conceptual differences, from Schaffer (2016: 79).

(with only the former taken to satisfy IF).¹¹ Wilson (2016) is clear that we will have to consider not only counterfactual scenarios, but sometimes also counterpossible ones, when we move to the grounding case. We may for example need to intervene on the existence of the singleton set without intervening on the existence of its member or intervene to change the value of $P \vee Q$ from true to false without changing the value of P from true to false. This will clearly violate IF. Wilson (2016) diagnoses the need for non-trivial counterpossibles as one of the reasons for grounding scepticism.¹²

While I do not think that the difficulties in understanding counterpossibles are trivial, I want to focus on a different problem in this chapter.¹³ Namely, even if we allow that there are non-trivial counterpossibles, there are stark differences in how we can understand the *aptness* of causal versus grounding models. That is, there are important differences in how we can judge whether or not a model is any good as a representation of a scenario. I turn to this question in section 3.

3. Aptness of Structural Equations Models

Once we have a causal model or a grounding model on the table the obvious question to ask is whether the model is an adequate representation of the causal or grounding scenario that we are modelling. For short, is the model an apt one? In this section I move on to consider what we can say about how we go about judging whether a model is an apt one. I will first show how the question of aptness arises and discuss how it is addressed in the context of causal structural equations models. In section 3.2 I will argue that the question of models' aptness is of similar importance in the grounding case, but that we do not have the same resources to address the question.

3.1 *Aptness of causal structural equations models*

Within the literature on causal modelling it is acknowledged that it is of crucial importance to select appropriate variables.¹⁴ We have already seen one suggested restriction on variable choice for causal models; when representing relations of causal dependence, the variables used should be independently fixable. Here, I want to focus on another aptness restriction: the appropriate grain of representation of the model. A model can fail to be apt by including too few variables or by including too many.

¹¹ Woodward's (2015) interest is in models with mixed causal and non-causal dependencies. Here he suggests respecting definitional constraints so that we should not "keep fixed" values of variables related in a definitional way to a variable under intervention.

¹² Schaffer (2016: 72) also takes this to open the door to grounding scepticism (although he does not endorse it).

¹³ I do not want to suggest that these are the only difficulties with extending the interventionist framework to grounding cases. See Koslicki (2016) for some very different concerns from the ones that I will raise here.

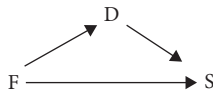
¹⁴ For an extended discussion about variable choice see, for example, Hitchcock (2012) and Woodward (2016).

Let me extend the example of the falling boulder from section 2 to illustrate the idea that it is possible to fail to construct an apt model by including *too many* variables on a path. In the new scenario, we add the information that the hiker survives the encounter with the falling boulder. The example is described by Hitchcock (2001: 276) and attributed to an early draft of Hall (2004).

“Boulder”: a boulder is dislodged, and begins rolling ominously toward Hiker. Before it reaches him, Hiker sees the boulder and ducks. The boulder sails harmlessly over his head with nary a centimeter to spare. Hiker survives his ordeal.

We can model the scenario by following Woodward’s (2003: 79–80) discussion of the case.

As before, we introduce a variable for whether or not the boulder falls, let us call it F , and a variable for whether or not the hiker ducks, let us call it D . The variables F and D behave as in section 2.1. We also introduce a variable for whether or not the hiker survives, let us call it S . S takes the value 1 if the hiker survives and 0 otherwise. In order to represent the relations of causal relevance we again make use of directed edges. Here, the fall (or non-fall) of the boulder is directly causally relevant to the survival or non-survival of the hiker. The fall (or non-fall) is also directly causally relevant to the ducking or non-ducking of the hiker. Finally, the ducking (or non-ducking) of the hiker is directly causally relevant to the survival or non-survival of the hiker. A redrawing of Woodward’s (2003: 79) figure 2.7.2 captures this.



We also need to represent the endogenous variables (S , D) as a function of their direct causes. As before $D=F$. We also need to express S as a function of its direct causes. This is easy to do. The hiker survives if either the boulder does not fall or the hiker ducks. So, the value of S is the highest of the values taken by either D or $1 - F$ (or both); $S = \text{Max}(D, 1 - F)$.¹⁵

Let us now return to the question of this section. Is the model an adequate representation of the scenario in question? In particular, have we included the right number of variables? Let us return to our intuitive judgements about the causal relations in the scenario. This case is often used to illustrate that at least some causal notions are not transitive. The boulder falling is a cause of the ducking, and the ducking is a cause of the survival of the hiker. Yet, the boulder falling is not a cause of the survival. The notion of cause that is relevant to our judgement is the notion of an actual cause (or token causation). Woodward’s (2003: 79–81) discussion of the case reveals that the result that our model delivers about whether or not the fall of the boulder is an actual

¹⁵ These are slight notational variants on Woodward’s (2003: 79) equations 2.7.4 and 2.7.5.

cause of the survival of the hiker hinges crucially on whether or not we judge it to be appropriate to include more variables along the F - S path in Figure 12.3.

Formally, for a case that is not one of symmetric overdetermination, we have the following criterion for whether $X = x$ is an actual cause of $Y = y$.¹⁶

(AC1) The actual value of $X = x$ and the actual value of $Y = y$.

(AC2) There is at least one route R from X to Y for which an intervention on X will change the value of Y , given that other direct causes Z_i of Y that are not on this route have been fixed at their actual values. (It is assumed that all direct causes of Y that are not on any route from X to Y remain at their actual values under the intervention on X .)

Then $X = x$ is an actual cause of $Y = y$ if and only if both conditions (AC1) and (AC2) are satisfied. (Woodward 2003: 77)

To assess whether $D=1$ is an actual cause of $S=1$, we evaluate the causal influence along each path from the ducking to the survival, keeping the off-path variables representing direct causes of S fixed to their actual values. If we keep the value of F (the falling of the boulder) fixed to its actual value (the only variable not on the path from D to S that is a direct cause of S) and we intervene to change whether the hiker ducks or not (the value of D), then we change whether or not the hiker survives (the value of S). So the ducking of the hiker is an actual cause of the hiker's survival. This is as we suspected.

However, the falling of the boulder is not an actual cause of the survival of the hiker. Following Woodward's discussion, we have two paths to consider. Let us look at the direct F - S path first. If we keep the ducking (D) fixed at its actual value, then changing whether or not the boulder falls (the value of F) does not alter whether or not the hiker survives (the value of S). We also have a second path to consider: the path from F to S that goes via D . Here there are no variables to keep fixed along the direct F to S path, so we cannot evaluate the influence of F on S via the F - D - S path by keeping any such variables fixed. In neither case do we find that the falling of the boulder makes a difference to the survival of the hiker. So far, things look good. Our use of structural equation modelling has recovered the judgement that the notion of actual cause fails to be transitive in the way that our intuitive judgement about the case leads us to expect.

Notice, however, that our solution depends on there not being a variable included on the F to S path representing a point on the trajectory where the boulder is too close to the hiker's head for the hiker to duck (as Woodward points out in his discussion of the case).

[T]his treatment of the boulder example depends crucially on the absence of any intermediate variable on the direct route from F to S . This raises the obvious question of why it wouldn't be equally or more correct to include such a variable in our representation of the example.

(Woodward 2003: 80)

If we introduce such a variable, then we need to keep this variable fixed to evaluate the influence along the F - D - S route. We are now considering a scenario where the boulder

¹⁶ We have to be more careful in cases of overdetermination. See footnote 17.

fails to fall but appears at an intermediate point in its trajectory where it is too late for the hiker to duck. If there was such a variable, then when evaluating the influence of *F* on *S* via the *F*–*D*–*S* path, we find that had the boulder not fallen the hiker would *not* have survived. What makes it inappropriate to include such a variable?

In our discussion of the falling boulder example... we rejected the idea that it was appropriate, given the causal structure of this example, to consider the possibility that the boulder both failed to fall and yet (somehow) appeared a few meters from the hiker's head. It was not that this was (in itself) a logical or causal or nomological impossibility, but rather that, to take this possibility seriously, we needed to consider an example with a rather different causal structure from the one we originally set out to analyze, one in which some independent mechanism or process, other than falling, is responsible for the appearance of the boulder in close proximity to the hiker. At least in ordinary contexts, the possibility that the boulder both fails to fall *and* appears near the hiker's head *and* doesn't get there as a result of following a trajectory from some independent source but instead, say, simply materializes near the hiker's head is not one that we are prepared to take seriously. (Woodward, 2003: 86)

The judgement that there is no serious possibility of the boulder failing to fall and yet simply materializing near the hiker's head is one that we have good evidence for. Although it is not ruled out by the metaphysical nature of causation or by the very concept of causation that a boulder could fail to fall and simply materialize on a trajectory close to the hiker's head, it is not how we take boulders to behave. The possibility considered is incompatible with our best theoretical understanding of the causal behaviour of objects like boulders. This is not to say that we take it to be incompatible with the *concept* of causation to have a boulder behave such as to simply appear close to the hiker's head. Rather our *a posteriori theory* of what the causal mechanisms and processes responsible for boulder behaviour are like rules out such a possibility. Like all of our theories about the physical world, such a theory is fallible and not typically a conceptual truth. For example, we may have a *theory* that restricts causal influences to propagate at subluminal speeds without taking it to be any part of the *concept* of causation that causal influences are so restricted.

The theory of the causal mechanisms involved is a *local* one. We can have a theory of the causal mechanisms behind boulder behaviour that rules out a boulder simply appearing close to the hiker's head without getting there by travelling in a continuous trajectory. Yet, we can also allow that an apt model for the behaviour of subatomic particles should not include a restriction where a particular kind of subatomic particle can only be present within a given area through the earlier presence of a particle of that kind at some nearby area. There is no conflict between these two judgements of aptness. They both stem from an understanding of what the causal mechanisms in question are—typically, and around here—like.

The considerations that we invoke in order to conclude that a model that includes an extra variable on the *F* to *S* path is not a good representation of the causal situation are knowable only *a posteriori*, they apply only to a particular type of system,

and they do not amount to a conceptual ban on the possibilities that are ruled out. In section 3.2 I will argue that this type of consideration is not available in the grounding case.

Before moving on to section 3.2 I want to push a bit deeper into the reasons that Woodward gives for ruling out the possibility of the boulder failing to fall but simply materializing close to the head of the hiker. Woodward takes this particular case to be one where the possibility is ruled out based on objective facts about how the world operates.

As I have already intimated, I think that it is true that in some cases an investigator's... interests and purposes (and not just how the world is) influence the possibilities that are taken seriously... On the other hand, as the examples described above illustrate, at least some of the considerations that go into such decisions are based on facts about how the world operates that seem perfectly objective. For example, there is nothing arbitrary or subjective about the claim that boulders don't materialize out of thin air... (Woodward 2003: 89)

The causal notions in the interventionist framework are not, however, generally purely a matter of just what the world is like.

[C]ausal judgments reflect both objective patterns of counterfactual dependence and which possibilities are taken seriously; they convey or summarize information about patterns of counterfactual dependence among those possibilities we are willing to take seriously. In other words, to the extent that subjectivity or interest relativity enters into causal judgments, it enters because it influences our judgements about which possibilities are to be taken seriously.

(Woodward 2003: 90)

To make the non-actual (but causally conceptually possible) scenario of the boulder materializing out of thin air ruled out on grounds of how the world operates, we need to appeal to some feature of the world that constrains not just what is *actually* true about the world; after all, we are ruling out a non-actual possibility. We need to turn to some feature of the world with modal force to achieve this. To introduce the notions of causal mechanism or causal production (or, for that matter, worldly dependence) is attractive in the boulder case since those notions are typically taken to be at least candidates for objective, interest independent, features of the world that do have modal force. A purely worldly notion of cause is also not, however, the notion of cause that is generally represented by causal graphs and structural equations in the interventionist account.

So far, I have tried to convince you that adjudicating whether or not a causal model includes the right variables (both in number and kind) to represent the causal structure of interest can involve considerations that, first, take us outside the realm of a purely conceptual account of causation and to a theory of features of particular causal processes or mechanisms. Crucially, we can have a posteriori evidence for or against a suggested theory of a particular causal process or a causal mechanism. Second, the theories of causal mechanisms or processes can be local. They do not have to be taken to apply to all types of causal processes. Third, in order to take the constraint on aptness

to stem from objective consideration from the way the world operates we need to introduce a notion of cause (or other notion with modal force) that is not the one that is the immediate target of analysis in interventionist causal models.

Let us see if we can generalize this discussion to help us answer the question of aptness in the case of grounding models.

3.2 *Aptness of grounding structural equations models*

The question of aptness is as important in the grounding case as in the causal case. In particular, the selection of which variables to include affects the conclusions about ground that can be drawn from the models. As the example of the boulder from section 3.1 illustrates, whether a model in the causal case delivers the intuitively correct verdict about actual causes depends on not including too many variables on a path. In this section I will show that including too many variables in a grounding model similarly leads to intuitively mistaken verdicts about the grounding relations in the scenario that is being modelled.

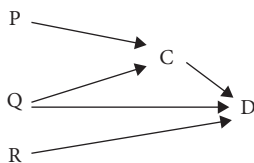
In the causal case this challenge was addressed by appealing to the aptness (or lack of aptness) of the model. At the end of this section I will argue that the kinds of considerations available in the causal case are not available in the grounding case.

Let me now turn to the example that will show that including too many variables is as problematic in the grounding case as in the causal case. Consider the following claim.

Disjunction from Conjunction (DfC): $P \wedge Q$ and R are both potential grounds for $Q \vee R$.

It is uncontroversial among grounding theorists that when R holds, R grounds $Q \vee R$ (it is similarly uncontroversial that Q can ground $Q \vee R$). It is less clear whether $P \wedge Q$ can ground $Q \vee R$. In favour of taking $P \wedge Q$ to be directly grounding relevant, we can note that it is easy to imagine scenarios where it being a fact that $P \wedge Q$ looks like a potential explanation of it being a fact that $Q \vee R$. For example, imagine a scenario where we start from information about the conjunctive fact that $P \wedge Q$; a tempting answer to give as to why $Q \vee R$ holds is that $P \wedge Q$ does. In particular, the extended interventionist motivation for attributing a relation of direct grounding relevance seems to hold: there are some interventions on the variable representing $P \wedge Q$ such that it is associated with changes in the variable representing $Q \vee R$ (in particular intervening in order to make the case that $P \wedge Q$ holds will make it the case that $Q \vee R$ holds, and some interventions to make it the case that $P \wedge Q$ does not hold will make it the case that $Q \vee R$ does not hold). Moreover, we cannot break the grounding structure into two steps; we cannot conceive of the scenario as one where $P \wedge Q$ grounds Q which grounds $Q \vee R$. Any standard account of ground would deny that $P \wedge Q$ is grounding relevant to Q . If we want to capture a grounding scenario where our “initial” conditions (fundamental conditions relative to the scenario modelled) are conjunctive facts then we have to either fail to capture any grounding relevance between $P \wedge Q$ and $Q \vee R$ or include it directly.

Whichever way we go on the truth or falsity of DfC , the grounding model below should strike us as problematic.



Here C is a variable that takes the value 1 if $P \wedge Q$ holds and 0 otherwise. D is a variable that takes the value 1 if $Q \vee R$ holds and 0 otherwise. P is a variable that takes the value 1 if P holds and 0 otherwise (and mutatis mutandis for Q and R). We can use the following structural equations to try to represent these relationships. The value of D is the highest value taken by any of R, C, Q ; $D = \text{Max}(R, C, Q)$. The value of C is the lowest value taken by either P or Q (or both); $C = \text{Min}(P, Q)$.

Let it be the case that $P, Q, P \wedge Q$ and $Q \vee R$ all obtain. Let it also be the case that R does not. In this scenario, Q should turn out to ground $Q \vee R$ but R should not. We can extend the terminology used earlier to capture this. In particular, $Q=1$ should turn out to be an actual ground of $D=1$ but $R=0$ should not turn out to be an actual ground of $D=1$. We have a few options available to us for how we go about evaluating the relations of actual ground in this case (corresponding to various alternatives for how to evaluate actual causation). Let us try a fairly straightforward and standard proposal by generalizing Woodward's criterion for cases that do not involve symmetric overdetermination. This is appropriate since, intuitively, in the scenario considered $Q \vee R$ is not actually overdetermined; Q holds but R does not. We will try to isolate the grounding influence along a single path. To do so we will keep any off-path variables representing direct grounds of D fixed at their actual values. If we follow this reasoning, then Q is not an actual ground of $Q \vee R$. Why? There are two paths from Q to D to consider: the direct path and the path via C . Let us consider the direct Q to D path first. There are two variables to keep fixed at their actual values: C and R . The actual value of C was stipulated to be 1. So changing the value of Q from 1 to 0 will not change the value of D . Let us consider the Q to C to D path. Now, we cannot fix the values of intermediate variables along the direct Q to D path (since there are no such variables). So we cannot evaluate the influence along the Q to C to D path. In analogy to the boulder case, $Q=1$ turns out to not be an actual ground of $D=1$. Here, however, this is bad news! Q should have turned out to be an actual ground for $Q \vee R$ in this scenario. The model has clearly gone wrong somewhere. The crucial question for grounding theorists is to identify where it has gone wrong.¹⁷

¹⁷ In making this evaluation I have relied on Woodward's (2003: 77) criteria for actual causation. One possible objection is that Woodward already notes that these criteria will have to be amended in order to handle cases of symmetric overdetermination well. However, this will not help the situation. On more involved definitions of actual causation adapted for the grounding case we can get Q to be an actual ground

Intuitively it is clear what has happened. While it is at least not clear that we would want to deny that the fact that $P \wedge Q$ could be grounding relevant to the fact that $Q \vee R$, it is a mistake to represent the fact that $P \wedge Q$ as an intermediate step between Q and $Q \vee R$ in the grounding structure in question. Just as in the discussion of including an intermediate variable on the F to S path in the boulder case, the mistake lies in including an intermediate variable where there ought to be none. In the boulder case we diagnosed the mistake by appealing to considerations from what we take the causal processes and causal mechanisms of boulder motion to be like. Can we do something similar in the grounding case?¹⁸ I think that the answer is that we cannot.

In the causal case we relied on a posteriori evidence for the claim that the causal mechanisms and processes involved in boulder motion operate to produce contiguous motion. In our experience, boulders (and boulder-like objects) do not just materialize out of thin air. Importantly, this is merely a theory about the causal behaviour of boulders. It does not rise to the level of a conceptual claim about causation.

We have several reasons to worry about whether something similar could apply to the grounding case. First, can we make use of questions about the relevant grounding mechanisms and processes when evaluating the aptness of a grounding model? I take it to be much easier to see how we have explanatory dependence in the examples of grounding in section 2 than to see how we have a relation of grounding mechanism or some specific grounding process.¹⁹ This would already seem to indicate a difference from the causal case; to get something close to the causal notion of process we would have to take notions of metaphysical building or metaphysical making non-metaphorically and to not be mere alternative ways of talking about metaphysical explanation and dependence.

Second, in the causal case we are relying on local theories of the causal behaviour of a specific type of system. We do not need to judge it to be conceptually or metaphysically impossible that boulders would materialize out of thin air in order to rule out a causal model as failing to be apt if it reflects such a possibility. In the grounding case this middle ground option—between including all possible scenarios and only including only one—is not available. What we would need is some information about the way that the actual world operates that would make it inappropriate to consider the possibility represented by including the variable representing the value of $P \wedge Q$, but that would not rule out the possibility that in other possible (or impossible) worlds the inclusion of such a variable would be appropriate. The relations that we are interested

for $Q \vee R$. However, we do so at the cost of allowing $P \wedge Q$ to count as a *separate* actual ground for $Q \vee R$. That is, we make the situation look like a case of actual overdetermination. This is not an improvement. When Q holds and R does not hold, $Q \vee R$ is not actually grounding overdetermined. See Weslake (forthcoming) for a summary of different proposals and for a new suggestion.

¹⁸ The criteria for aptness of grounding models that Schaffer (2016: 74–5) proposes (by adapting to the grounding case the criteria suggested by Blanchard and Schaffer 2017 for the causal case) do not solve the problem. The problem is not that we have *left out* some fact that ought to have been included (we do not have too few variables).

¹⁹ This is in line with Wilson (forthcoming). However, Schaffer (2016: 54) takes ground to have most of the features that we may associate with causal relations.

in when it comes to grounding cases do not, in general, seem to allow for this. We do not have a theory of disjunction in the actual world that—with good reason—takes disjunction to have features at the actual world that are different from those of disjunction in other possible worlds (although we can, of course consider different types of disjunction). A local theory of disjunction is not forthcoming.

In the causal case, it is the appeal to such local considerations that allow us to put objective constraints on which models are apt. Without them we are left to select the possibilities that are to be taken seriously merely based on criteria such as the interest of the modeller and other non-objective factors. In the grounding case we need to find a substitute for local theories of causal mechanisms or processes that could play the role of selecting the relevant possibilities (or impossibilities) that ought to be captured by the grounding model.

The most obvious solution to the problem—take all possibilities seriously in the grounding case—will not work. After all, we have already seen that the grounding models in question rely on taking some *counterpossibles* seriously. We cannot restrict ourselves to merely taking all possibilities (not even all logical possibilities) seriously. However, merely including impossibilities does not help. If we extend the proposal to include counterpossibles—take all possibilities and impossibilities seriously in the grounding case—we destroy the ability to rule out the graph in Figure 12.4 as failing to be apt. There is no scenario (possible or impossible) that the graph in Figure 12.4 could mistakenly have represented as a scenario that ought to be taken seriously. They should all be taken seriously on the view under consideration.

So far I have hoped to convince you that when it comes to considerations of aptness the analogy between causal models and grounding models break down. In section 4 I will briefly suggest a way forward for grounding models.

4. Conclusion

Grounding is often taken to have very close ties to metaphysical explanation. The existence of a type of non-causal explanation has been cited as the driving reason to postulate a relation of ground. Here, I have simply granted that there are such explanations and focused on the question of whether we have reason to take ground to be akin to causation in its role in explanation. I think that the answer to this question is no.

Earlier I argued that we do not have an obvious way of extending to the grounding case the resources that we have for adjudicating whether or not a causal model is apt in the causal case. The difficulty in the grounding case of finding reasons for ruling out a possibility (or counterpossibility) as not appropriate also suggests a way forward. We cannot simply include all possibilities or all possibilities and counterpossibilities. Nor can we appeal to a posteriori evidence for local theories of grounding processes. We can, however, appeal to general logical and metaphysical theoretical principles to try to provide an account of which possibilities or counterpossibilities are relevant.

The focus on exceptionless privileged regularities puts the account more in line with the deductive-nomological account than with causal accounts of explanation.²⁰ Part of what is distinctive about causal explanation and causal modelling is that we can use it even when we lack information about exceptionless laws. In the causal case structural equations do not need to be exceptionless, and our judgements about the aptness of models can rely on empirical local theories of particular causal processes. Earlier I argued that the same does not hold for the case of grounding models.

If the explanations in the grounding case are better understood by analogy to explanations that invoke general laws or general principles than by analogy to causal explanations, then the questions that we raise about ground look rather different. It is now a pressing matter to understand these general principles. They are what we need to articulate in order to substantiate a grounding claim.

Although my argument here has been different from Wilson's (2014: 544–5), I take my argument to support the claim that bare "... Grounding... claims leave open questions that must be answered to gain even basic illumination about or allow even basic assessment of claims of metaphysical dependence". In particular, we cannot avoid answering questions about the aptness of grounding models. Moreover, the answers look like they will—unlike in the causal case—have to come from general principles.

Acknowledgements

Thank you to the Nottingham Metaphysics and Epistemology Reading Group, the editors of this volume, and to an anonymous referee for comments on an earlier version of this chapter. Thank you also to Jonathan Tallant and Al Wilson.

References

- Audi, P. (2012), 'A Clarification and Defense of the Notion of Grounding', in F. Correia and B. Schnieder (eds.), *Metaphysical Grounding: Understanding the Structure of Reality* (Cambridge: Cambridge University Press), 101–21.
- Blanchard, T. and Schaffer, J. (2017), 'Cause without Default', in H. Beebe, C. Hitchcock, and H. Price (eds.), *Making a Difference: Essays on the Philosophy of Causation* (Oxford: Oxford University Press), 175–214.
- Bliss, R. and Trogon, K. (2014), 'Metaphysical Grounding', *The Stanford Encyclopedia of Philosophy* (Winter 2014 Edition), Edward N. Zalta (ed.). <<http://plato.stanford.edu/archives/win2014/entries/grounding/>>.
- Dasgupta, S. (2014), 'On the Plurality of Grounds', *Philosophers' Imprint* 14: 1–28.
- Hall, N. (2004), 'Two Concepts of Causation', in J. Collins, N. Hall, and L. A. Paul (eds.), *Causation and Counterfactuals* (Cambridge, MA: MIT Press), 225–76.

²⁰ For an explicit articulation of ground by analogy to law-based explanation see Wilsch (2015, 2016) and Jansson (2017).

- Hitchcock, C. (2001), 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98: 273–99.
- Hitchcock, C. (2012), 'Events and Times: A Case Study in Means–Ends Metaphysics', *Philosophical Studies* 160: 79–96.
- Illari, P. (2013), 'Mechanistic Explanation: Integrating the Ontic and Epistemic', *Erkenntnis* 78 supplement: 237–55.
- Jansson, L. (2017), 'Explanatory Asymmetries, Ground, and Ontological Dependence', *Erkenntnis* 82: 17–44.
- Jenkins, C. S. I. (2013), 'Explanation and Fundamentality', in M. Hoeltje, B. Schnieder, and A. Steinberg (eds.), *Varieties of Dependence* (Munich: Philosophia Verlag), 211–42.
- Kim, J. (1974), 'Noncausal Connections', *Noûs* 8: 41–52.
- Koslicki, K. (2016), 'Where Grounding and Causation Part Ways: Comments on Schaffer', *Philosophical Studies* 173: 101–12.
- Kovacs, D. M. (2017), 'Grounding and the Argument from Explanatoriness', *Philosophical Studies* 12: 2927–52.
- Schaffer, J. (2016), 'Grounding in the Image of Causation', *Philosophical Studies* 173: 49–100.
- Weslake, B. (forthcoming), 'A Partial Theory of Actual Causation', *British Journal for the Philosophy of Science*.
- Wilsch, T. (2015), 'The Nomological Account of Ground', *Philosophical Studies* 172: 3293–312.
- Wilsch, T. (2016), 'The Deductive-Nomological Account of Metaphysical Explanation', *Australasian Journal of Philosophy* 94: 1–23.
- Wilson, A. (2016), 'Grounding Entails Counterpossible Non-Triviality', *Philosophy and Phenomenological Research*, online first. DOI: 10.1111/phpr.12305.
- Wilson, A. (forthcoming), 'Metaphysical Causation', *Noûs*.
- Wilson, J. (2014), 'No Work for a Theory of Grounding', *Inquiry: An Interdisciplinary Journal of Philosophy* 57: 535–79.
- Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation* (New York: Oxford University Press).
- Woodward, J. (2015), 'Interventionism and Causal Exclusion', *Philosophy and Phenomenological Research* 91: 303–47.
- Woodward, J. (2016), 'The Problem of Variable Choice', *Synthese* 193: 1047–72.

Index

- “3M” constraint on explanation 146, 168–70
- abstract explanations 41–5, 47, 51, 53, 64–5
- action explanation 4n
- aeolian geomorphology 142–3, 149
- aetiological explanation 166–7
- aims of science 1, 66–7
- Andersen, H. 46n, 47, 160, 166
- Anderson, R. 150–4
- asymmetry of explanation 92–3, 188, 199–200
- Audi, P. 252
- Ayrton, H. 141
- baby-carriage example 16, 18–19, 25–6, 28, 30
- Bagnold, R. 151, 158
- Baker, A. 209, 231n
- Barlow, H. 173, 177–8
- barometer example 47–8
- Barr, M. 244, 238n, 244
- Batterman, R. 63–5, 87, 132, 134–5, 145–8, 160, 206, 211
- Blanchard, T. 251
- Bokulich, A. 66, 68, 122
- Bondi, H. 16
- Boyle’s law 98, 243
- Brading, K. 188
- Brigandt, I. 41n
- bridges of Königsberg example 15, 19, 32–3, 45, 83–4, 90–1, 98–9, 100, 109–13, 127–8, 203
- Brown, H. 20n, 200
- Bueno, O. 100
- Callender, C. 36n, 125
- Castellani, E. 188
- causal imperialism 141
- causal laws 16, 18, 24, 36, 76, 104, 115, 147, 160
- causal mechanism 2, 260
- causal anti-foundationalism, *see* causation, anti-foundationalism
- causal process 2, 17, 30, 58, 99, 104–12, 143, 149, 151–60, 166, 168, 260–4
- causal reductionism 5, 7, 9, 75, 77, 96, 115
- causal theory of explanation 2, 43, 57, 74, 96
- causation,
 - anti-foundationalism 3–4
 - causal difference-making 102–6, 108
 - causal laws 16
 - concept of 259
 - counterfactual theory of causation 30–1, 80, 104, 119–22
 - causal difference-making 102–6, 108
 - dispositionalism 191, 193, 201
 - folk theory 148n
 - interventionist theory of causation 23–4, 78–80, 119–24, 164–5, 175n, 251–65
 - theories of causation in general 2
- central limit theorem 215, 219, 225
- cicadas example 141n, 209
- chaos 223–4
- Colyvan, M. 100, 243
- communicative accounts 59–61, 63, 66, 69, 71
- conservation laws 16–20, 24, 26–7, 187, 197–203
- constitutive explanations 41–53
- contrastive accounts of explanation 50–2, 59–60
- contrastive facts and causal explanation 50–5
- constraints (*see also* explanation by constraint) 17–18, 24–8
- contrastive facts and non-causal explanation 53
- counterfactual theory of explanation 23, 46, 58, 78–9, 119, 122–4, 186–202
- counterpossibles 169, 175, 254n, 256, 264
- covering law model of explanation (*see also* deductive-nomological model) 1, 47, 57, 64, 57, 77, 81–2, 143, 179
- Craver, C. 69, 142, 146–8, 168
- critical phenomena 211
- Dasgupta, S. 250
- deductive-nomological (DN) model (*see also* covering law model) 20n, 61, 188, 265
- defect dynamics explanation (of sand ripples) 150, 155–9, 160–1
- dependence, explanatory, *see* explanatory dependence
- dimensional explanations 36
- dimensionality of space example 22, 35, 36n, 74–5, 87, 123–5, 128, 200
- dispositionalist theory of causation, *see* causation
- distinctively mathematical explanation (*see also* mathematical explanation in science) 201, 210
- double pendulum 114–15, 210, 218
- Douglas, H. 66
- Earman, J. 21
- eikonic conception of explanation 142–4
- Einstein, A. 21, 54
- epistemic accounts of explanation 39–42

- equilibrium explanation 110–11, 114, 130, 132, 208
 Euler, L. 83–4, 98
 Euler's explanation, *see* bridges of Königsberg example
 Euler's theorem 15, 83–4, 98–9, 127
 explanation,
 in logic 4
 in mathematics 4, 23, 231–47
 in metaphysics 4, 251–65
 explanatory asymmetry, *see* asymmetry of explanation
 explanation by constraint 17–37, 114–15
 explanatorily derivative laws (EDLs) 27, 33–6
 explanatorily fundamental laws (FDLs) 27, 33–6
 explanatory dependence 57, 58, 60, 63, 65, 67–9, 185, 194, 201, 263
 explanatory irrelevance 30–1, 111, 128–36
 explanatory monism 5–6, 77–8
 explanatory pluralism 1, 5, 36–7, 39–42, 49, 55, 75–7, 142–4, 161
 explanatory proofs 231–47
 explanatory virtues 39, 232, 233n, 242–7
 extra-mathematical explanation,
 see mathematical explanation in science

 Feigenbaum, M. 223–4
 Fermat's explanation 15, 19, 32–3, 82–3, 90
 Feynman, R. 202
 Fine, K. 48
 free group theorem 234–42, 245, 246
 Friedman, M. 57n, 186

 gas pressure 97–8, 108–9
 graphs 83, 118, 251, 253
 group axioms 234
 group theory 187, 189, 232–42
 Goldenfeld, N. 134, 145, 225
 grounding (explanation) 9, 19n, 75, 250–65

 Hall, E. 257
 Hamilton walk 109–10
 Haslanger, S. 45n
 Hempel, C. 1, 61, 77, 79, 81–2, 89–90, 178
 Hempel's monism 77, 79
 Hertz, H. 18–19, 22–3, 35–6
 Hildreth, E. 172–3
 Hitchcock, C. 50n, 257
 Holland, P. 200
 how-actually vs. how-possibly explanations 80
 Humean approach (to laws) 201–2
 Huneman, P. 118

 Idealizations 6, 41, 62–6, 80, 82n, 86, 144, 153, 225
 information theory 175
 intra-mathematical explanation 232

 irrelevance, *see* explanatory irrelevance
 Ising model 214–15, 220–2, 224

 Jansson, L. 146
 Jenkins, C. 250

 Kadanoff, L. 134, 145, 212n, 214n, 225
 Kahneman, D. 87–8
 Kahnemann and Tversky's explanation 87–8, 91–2
 kairetic account of explanation 63, 102–7, 194
 Kaplan D. 146, 148, 165n, 166, 168, 170
 Khinchin, A. 219
 Kitcher, P. 186, 244, 245n
 Kocurek, G. 150, 154–7
 Koslicki, K. 48
 Königsberg, *see* bridges of Königsberg example

 Lange, M. 23, 26–7, 34, 37, 76–7, 114–15, 126, 141, 147, 159n, 160, 169, 185n, 186, 197, 202–3, 209–11, 218, 231n, 232n
 lateral inhibition 171
 lattice gas automaton (LGA) example 134–6
 Laughlin, S. 171, 177–9
 laws of association 75, 81
 laws of coexistence 75, 77, 81, 89–90, 190n
 levels of explanation 69–70
 Levy, A. 63
 Lewis, D. 30–1, 60, 69, 75–6, 104, 141n, 191
 Lipton, P. 39, 52–3, 60–1, 87
 logistic map 223
 Lombrozo, T. 62
 Lorentz transformations 18–23, 26, 28, 33–5

 Mach, E. 3, 77n, 171
 Mac Lane, S. 245
 Marr, D. 172
 mathematical explanation in science 41, 98–101, 114–15, 147, 159n, 166, 206, 209–10, 225, 232
 Maxwell, J. C. 153
 mean field theory (MFT) 215–16
 mechanistic explanation 8, 74, 167–72, 252n
 meta-laws 197, 202
 method of arbitrary functions 132–3
 minimal models 120, 133–5, 145–6, 148
 modal account of explanation 50, 76
 model-based explanation 144
 momentum conservation, *see* conservation laws
 monism, *see* explanatory monism
 multiple realizability 132n, 217

 necessity, *see* varieties of necessity
 networks 83, 118
 Noether's theorem 197–203
 northern elephant seals 97, 108, 110
 Norton, J. 66

- Oppenheim, P. 1
- ontic accounts of explanation 39–42, 46, 50, 52, 54, 59–62, 67–8, 69–72, 142–3
- ontological dependence 48–9
- Pauli exclusion principle 190–6
- pendulum explanation 81, 89–90
- Pexton, M. 145
- Pincock, C. 98, 100–1
- phase transition 211–14, 217n
- pluralism, *see* explanatory pluralism
- power law 214n
- pragmatic theory of explanation 3, 55, 59–60
- prediction 66–7, 71, 165, 176, 212, 215
- predictive coding 175–7
- principle of relativity 21, 23, 25–6, 28, 33–4
- probability theory 97, 207, 219
- psychology of explanation 7, 60, 62, 65, 71
- Putnam, H. 70
- quantum field theory 117, 212n, 218n
- quantum mechanics 141n, 144–5, 189–96, 215
- Railton, P. 105
- random genetic drift 97, 108
- Rayleigh–Bénard convection 96–7
- reductionism 5, 7, 9, 69–70, 75, 77, 155
- reductive explanation 69–70
- renormalization group (RG) explanation 84–7, 124, 134–6, 145–8, 207–25
- representation and explanation 62–70
- reptation model explanation (of sand ripples) 150–61
- relativity theory 18–19, 20–3, 26, 33–5
- relevance logic 243n
- representationalism (in mathematics) 99–101, 107, 115
- Reutlinger, A. 39n, 48n, 78–9, 124, 136, 145–8, 160, 206, 211, 221
- Rice, C. 132, 134–5, 145–6, 160
- Ross, L. 148, 160
- Russell, B. 3–4, 89
- Russellian strategy 88–9
- Russellian criteria 89, 148n
- Saatsi, J. 145–6, 194n
- Salmon, W. 1, 17, 34, 35, 58, 76
- Schaffer, J. 250–1, 254, 263n
- self-organization 155
- semiclassical 144, 145n, 160n
- Skow, B. 75–6, 141n, 191
- Sober, E. 110–11
- statistical explanation 36–7, 74, 87–8, 91, 111
- statistical mechanics 8, 85–6, 207–8, 211, 215, 219–20, 224, 243
- Steiner, M. 23n, 243, 243n
- Sterling, P. 177–9, 182
- strawberries example 15, 19, 31–2, 119, 126–7, 147
- Strevens, M. 61, 63–5, 110–11, 146, 194, 244n
- structural equation (model) 250–65
- structural model explanation 144–5
- symmetrization postulate 189–92, 195
- symmetry breaking 212, 222
- symmetry principles 28, 74, 185n, 186, 202, 210
- teleological explanation 4n
- Tversky, A. 87–8
- understanding (in science) 60–2, 71, 106–8, 111, 119,
- unification 244–5
- unificationist theory of explanation 3, 57, 68, 143, 178, 186
- unity of science 69–70
- universal phenomena 85–7, 91, 124, 134–5, 146–7, 160–1, 196, 207, 211–18
- universal property 234, 242–3
- universality class 85–7, 91, 124, 134–5, 146–7, 211–15, 221–2
- Van Fraassen, B. 60, 69
- variational principles 82
- varieties of necessity 16–17, 24–8, 89, 91–2
- Watkins, J. 34n
- wavefunction scarring 144, 160n
- Werner, B. 150, 155–7
- Wigner, E. 16–17
- Wilson, A. 250, 254, 256, 265
- Wilson, K. 212, 214, 219
- Woodward, J. 18, 23–4, 43–4, 46, 55, 78–9, 80–1, 144, 164, 185, 188, 251, 253, 255–62
- Wright, C. 54n