OXFORD

LEVELS OF EXPLANATION

Edited by Katie Robertson & Alastair Wilson

Levels of Explanation

Levels of Explanation

Edited by

KATIE ROBERTSON
AND
ALASTAIR WILSON

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

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: 2024939406

ISBN 978–0–19–286294–5

DOI: 10.1093/oso/9780192862945.001.0001

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

Contents

PART I: FOUNDATIONS OF EXPLANATORY LEVELS

PART II: LEVELS OF EXPLANATION IN CAUSAL MODELLING

PART III: LEVELS OF EXPLANATION IN HIGHER-LEVEL SCIENCES

PART IV: LEVELS OF EXPLANATION IN PHYSICS

PART V: LEVELS OF EXPLANATION IN MATHEMATICS AND METAPHYSICS

PART VI: HOW ARE EXPLANATORY LEVELS POSSIBLE?

Acknowledgements

We are extremely grateful to the many people who have helped with the genesis and production of this volume. First to our contributors—thank you all for being essential parts of the project, and for your patience while it came together. Next, thanks to our colleagues in FraMEPhys—especially John Barnden, Jean Baynham, Konner Childers, Nicholas Emmerson, Joaquim Giannotti, Michael Townsen Hicks, Francis Longworth, Dan Marshall, Vera Matarese, John Murphy, Zee Perry, Josh Quirke, Noelia Iranzo Ribera, Aaron Sloman, and Nanxin Wei—for being an endless source of insight and solidarity. Thanks also to colleagues near and far who have helped in ways large and small—especially Sam Baron, Helen Beebee, Alexander Bird, Nikk Effingham, Salvatore Florio, Lina Jansson, Nicholas Jones, Iain Law, Baptiste Le Bihan, Heather Logue, Niels Martens, Martin Pickup, Michael Rush, Juha Saatsi, Maja Spener, Scott Sturgeon, Henry Taylor, and Robbie Williams. We are also grateful to Miranda Rose for proof-reading and to Niccolò Aimone Pisano for composing the index. Finally, our thanks go to the careful and constructive reviewers for Oxford University Press, and to Peter Momtchiloff for his confidence in the project.

This volume is part of the project A Framework for Metaphysical Explanation in Physics (FraMEPhys), which received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 757295).

Contributors

Carolin Antos, University of Konstanz

William Bechtel, University of California, San Diego

Harjit Bhogal, University of Maryland

Mazviita Chirimuuta, University of Edinburgh

Mark Colyvan, University of Sydney

Karen Crowther, University of Oslo

Nina Emery, Mount Holyoke College

Alexander Franklin, King's College London

Michael Townsen Hicks, University of Glasgow

Vera Hoffmann-Kolss, University of Bern

Harold Kincaid, University of Cape Town

Eleanor Knox, King's College London

Christian List, LMU Munich

Kerry McKenzie, University of California, San Diego

Angela Potochnik, University of Cincinnati

Katie Robertson, University of Stirling

Michael Strevens, New York University

Elanor Taylor, Johns Hopkins University

Brad Weslake, NYU Shanghai

Alastair Wilson, University of Leeds

David Yates, University of Lisbon

Introduction

Levels of Explanation

Katie Robertson and Alastair Wilson

The world around us can be explained on many different levels. 'Why'-questions can have multiple distinct correct answers, with context playing a central role in determining what is asked and how it is answered. Different explanations of the same phenomenon can enhance, rather than exclude, one another; a complete understanding of the phenomenon requires grasping all these explanations. Understanding levels of explanation and how they relate is the main project of this volume.

Explanations at different levels, distinctively, complement rather than clash. That is to say that distinct candidate explanations at a single level tend to exclude one another in a way in which candidate explanations at different levels need not. The bushfire was very likely started either by a lightning strike or by a discarded match; it is very unlikely that any individual fire was started by both acting together. But explanations of the bushfire in terms of a lightning strike and in terms of the effects of climate change on extreme weather events need not exclude each other; the former may explain the fire at the level of weather, while the latter explains the very same fire at the level of climate. Each explanation plays a role in our overall understanding of the fire.

Levels of explanation, as we understand them here, are commonplace and may be grasped from an early age. Children are voracious consumers of explanations. When a curious child asks a question, 'why P?' and receives an answer, 'because Q', typically she will accept the answer and move on, or ask a follow-up question, 'why Q?'. But sometimes she will frown, and repeat her initial question. It turns out that further elaboration of the answer Q does not help; she understands it perfectly well already. What her frown is calling for is an explanation on a different level from the one she has been offered.

Our practice of multilevel explanation is familiar and everyday, but it also takes centre stage in some of the most sophisticated contemporary work in philosophy of science and metaphysics. In this volume we have collected together some of the best of this philosophical work to give an overview of explanatory levels[1] and of

[1] We use 'explanatory level' and 'level of explanation' interchangeably.

Katie Robertson and Alastair Wilson, *Introduction* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Katie Robertson and Alastair Wilson 2024. DOI: 10.1093/oso/9780192862945.003.0001

their applications. Explanations at multiple levels can coexist without conflict or redundancy, can be employed for different epistemic or pragmatic purposes, and can be combined together to give richer explanatory models. These features give them broad potential application and so it is no surprise that talk of levels of explanation is commonplace across a wide range of disciplines. This book touches *inter alia* on explanatory level structures in cognitive science, sociology, molecular biology, materials science, particle physics, geometry, set theory, and the metaphysics of dispositions.

While many of the contributors to this volume are enthusiasts about levels of explanation—seeking to either contribute to our understanding of levels and their scientific usefulness, or to apply levels to solve philosophical problems raised by science, mathematics, or metaphysics—the concept has also come in for plenty of critical discussion. Explanatory level sceptics in one way or another argue that levels frameworks distort the scientific reasoning they are intended to illuminate. Eronen (2015), for example, argues that the work of explanatory levels is better assigned to specific notions of scale and composition, while Potochnik (2017) likewise argues that there is no useful generalized notion of explanatory level. Sceptical voices are represented in this volume: in particular, the chapters by Bechtel and Chirimuuta question straightforward realism about levels of explanation, while Potochnik argues we should reject the entire framework. Others, such as Franklin, address extant challenges for explanatory level frameworks. However, most of the contributors are positive about the usefulness of explanatory levels, seeking either to understand levels or to exploit them in their various projects. One of our aims with the volume is to give a rounded picture of what the concept of levels of explanation can do for philosophers, and thereby to give a clearer idea of the costs that would be involved in rejecting it as the sceptics urge.

As an initial step in systematizing our thinking about levels of explanation, we can model levels as classes of answers to explanatory questions—typically ‘why?’-questions but potentially also ‘how?’-questions. For some given explanatory question there will usually be many prospective answers, and each question-answer pair may then be grouped into classes unified by the equivalence relation ‘is at the same level as’. Of the prospective question-answer pairs at a given level, one (or maybe more) may be correct; but there is no presumption that every explanatory question has a correct answer at every level. For further insight into the nature of levels of explanation, we then need to ask after the nature of the same-level-as relation (or, perhaps more saliently in some circumstances, the different-level-as relation): what is it for two candidate explanations of some phenomenon to be at the same level? Which features make for explanatory stratification? Different philosophical accounts of levels of explanation, including those explored in some chapters of this volume, fill in the details in different ways. In particular, the chapters by List, Chirimuuta, Kincaid, Crowther, Knox, and Strevens provide illustratively different accounts of the nature of explanatory levels which have this core structure in common.

Over recent decades, it has gradually become orthodoxy in the philosophy of explanation that the best explanations need not reside at the fundamental level. In Potochnik's chapter, this thesis is named *explanatory anti-reductionism*. The higher-level explanations which higher-level sciences provide are central to the case for the irreducibility of the respective higher-level theories, and consequently to the case for the reality of the emergent entities that these theories describe. Explanations in higher-level terms are said to be more proportionate (Yablo 1992), deeper (Hitchcock and Woodward 2003; Strevens 2008), more abstract (Knox 2016), and/or more computationally tractable (Weisberg 2007) than explanations given in fundamental terms.

Explanations which reside wholly at a single level—whether they are provided by (putatively fundamental) classical field theories or by (evidently non-fundamental) evolutionary biology—remain conceptually relatively straightforward. But as well as mapping individual explanations at higher levels, scientific practice is rich in explanations drawing on multiple levels at once. Many of the hardest philosophical puzzles of explanation derive from the interplay of explanatory factors across multiple levels of description, as when the cooling of a cup of coffee is explained in terms of collective molecular motion, or when a cognitive process causes a bodily movement. In some cases lower-level events might directly influence higher-level events, such as when an atomic decay causes an explosive nuclear chain reaction, and sometimes higher-level events might directly influence lower-level events, such as my desire to type causing my fingers to move. Any account of levels of explanation must be able to do justice to the links between levels.

How are higher-level explanations connected to and constrained by other levels of explanation? Strategies for answering this question have tended to vary across subdisciplines. Debates about multilevel causation have a long pedigree in philosophy of mind, but recent work on causal modelling has both shed new light on mental causation and tied the problem closely to the philosophy of science, where a causal modelling account of explanation is dominant; Weslake's chapter provides an up-to-date account. More generally, enthusiasm about levels of explanation has tended to go hand in hand with enthusiasm about emergence in some domain or other. This motivating connection between emergence and explanatory levels plays a central role in the chapters by Knox, Franklin, and Crowther.

In philosophy of physics, theory reduction has often been the focus—at least since the highly influential work of Nagel (1935, 1949, 1961, 1970) (later refined by Schaffner (1967, 1990)) in which examples from physics played a central role. Reductionists in this tradition, including Lewis (1994), Loewer (2001), Butterfield (2011), and many others, have tended to endorse the idea of 'in principle' reducibility of high-level physical phenomena to lower-level phenomena. Meanwhile, anti-reductionists such as Cartwright (1999) and Dupré (1993) have argued we have no evidence for and some evidence against the thesis of in-principle-reducibility. However, authors including Batterman (2013, 2021) and Wilson (2017) have argued that this focus on reduction distorts scientific practice, arguing for a more nuanced picture. For Batterman, the 'tyranny of scales'—the stark choice between top-down and bottom-up explanatory strategies—is a false dichotomy.

In philosophy of biology, mechanisms have often been seen as key to understanding complex causal explanations. The right level at which to understand a biological system, in mechanistic terms, is one at which the system's behaviour can be understood as resulting from the interaction of distinct subsystems with distinct characteristic functions (Craver (2007), Bechtel (2008)). Another key concept has been levels of organization; see Wimsatt (1976) for an early statement, and Brooks, DiFrisco, and Wimsatt (2021) for the state of the art. Levels of organization bear some similarities to our general concept of levels of explanation, but also differ in certain respects. We discuss both mechanistic levels and levels of organization below, as potential rivals to the explanatory levels approach.

In metaphysics, levels of explanation have often been discussed in the context of Jaegwon Kim's causal exclusion argument (Kim 1983); levels offer a possible route to a robust explanatory role for higher-level properties in mental causation, free action, and related phenomena. Metaphysical relations including grounding and realization have been offered as candidates for the level-connection relation between scientific levels; grounding is often supposed to connect levels 'vertically', with causation connecting events 'horizontally' (Fine (2012), Bennett (2017)). But we can also identify cases of levels of explanation which are distinctively metaphysical: in many cases, theorizing in metaphysics aims to provide explanations at some level even more fundamental than the level of fundamental physics. When metaphysicians aim, for example, to account for the instantiation of quantitative properties in terms of relations to universals (Mundy (1987)) or in terms of mereology (Perry (forthcoming)), or when they aim to reduce spacetime to a causal structure between events (Baron and Le Bihan (forthcoming)), they can be regarded as operating at a distinctively metaphysical level of explanation.

The levels of explanation framework as we understand it here is general, in that it is applicable to many kinds of investigative context and to many varieties of explanation—not just to causal explanation, for example. This makes the framework flexible and open-ended, such that one can see different chapters of this volume as developing the general approach in very different ways. But the explanatory levels approach is not without distinctive content; not just any old description counts as an explanatory level. To provide a clearer sense of the levels concept we have in mind, we next contrast it with other approaches to levels which have appeared in the philosophy of science literature.

Rival one: descriptive levels. A descriptive level is a kind of imprecision in description: we describe our target system up to a certain level of detail, but not beyond. As authors including Lewis (1988a, 1988b) and List (2019) have emphasized, descriptive levels can be readily characterized using only simple modal resources: the basic idea is that the less-detailed/coarser-grained description supervenes on the more detailed/finer grained description, but not vice versa. However, for our purposes descriptive levels are too thin.[2] For a start, they are cheap; any arbitrary choice of details to exclude or include, however gerrymandered, characterizes a descriptive level. More tellingly, descriptive levels bear no direct epistemic significance; we learn nothing of any substance about a system when we are told that it can be described at a certain level of detail. Dropping decimal places in a quantitative description would, on this approach, automatically give a new level—without this being at all enlightening. Knox emphasizes this point, arguing that levels which are useful in physics need to be linked to *useful* variable changes rather than to any variable change whatsoever. Identifying distinct levels of explanation of some phenomenon is a non-trivial epistemic achievement which requires us to be able to identify—if not yet answer—distinct classes of explanatory questions concerning the phenomenon.

Rival two: levels of scale. Here talk of levels is understood as relating entirely to phenomena analysed at different physical scales—for example, at a length scale corresponding to metres or at an energy scale corresponding to gigaelectron-volts. This representation of levels is often found in introductory textbooks in biology or in introductory physics courses, where one presents the subject as spanning/investigating from the scale of subatomic particles to the scale of galaxy clusters. But there are also more sophisticated accounts of inter-theoretic relations which focus on scale: Ladyman and Ross (2007) incorporate the notion into their distinctive thesis of scale-relativity of ontology. Thinking of levels of explanation in terms of scale doesn't work across the board, however: variation in scale is too inflexible to accommodate all of the different ways in which explanatory levels might interrelate. For example, the explanatory levels structures characterized by List in Chapter 1 are typically partial orderings, while the ordering of levels on any given scale is a total ordering; Potochnik's chapter also discusses this point. Kincaid's chapter argues that scale is insufficient for understanding levels of explanation in social science. And Knox discusses the example of the 'level of Newtonian physics': this way of thinking is useful for some explanatory purposes, but it does not correspond neatly to any physical scale. In some cases perhaps—in particular the literature on effective field theories (Castellani (2000), Franklin (2018), Wallace (2019))—scale and explanatory levels do correlate well. But the link between scale and explanatory levels is not fully general.

[2] This is not to say that descriptive levels are useless, of course; they have been employed as modelling devices within metaphysics and philosophy of language (Lewis (1988a, 1988b), Yablo (2017)). Indeed, descriptive levels may be employed as part of a substantive theory of levels, as in List's approach in Chapter 1; our point here is that some further ingredient is needed.

Rival three: compositional levels. Here we have in mind the tradition following the classic paper of Oppenheim and Putnam (1958), who identify entity-based levels such that the smaller things at one level generically compose the larger things at the next level up. Other more sophisticated versions of this approach include Schaffer (2009), who uses grounding as an interlevel link as part of his priority monism. Compositional levels find few supporters these days, however, because of the way they tend to generate an exclusive focus on the compositional and mereological relations between the entities they structure. The 'Lego' view of science no longer seems tenable: interaction effects between the constituents of composites are endemic in real-world level structures. An atom isn't just a simple aggregate of its constituent nucleons and electrons; instead it derives nearly all of its interesting physical and chemical properties from the balance of interactions between these constituents. Likewise, in the classic application of explanatory levels to cognitive science by Marr (1981) discussed in Chirimuuta's chapter, Marr raises the worry that the whole can't be effectively studied by studying only the parts.

Rival four: levels of organization. Taking account of the limitations of compositional accounts of levels tends to take us towards a different class of rival approaches, which retain certain compositional aspects but emphasize organization of the elements. This category most prominently includes the kind of levels of organization deriving from the work of Wimsatt (1976), which are prominent in the philosophy of biology. Crowther's chapter also touches upon the usefulness of levels of organization in physics, and they can be modelled using List's notion of an ontological level. For our purposes, however, levels of organization are unsuitable: they lack a sufficiently direct link to explanatory value. The world may be organized in all sorts of different ways at different scales, but only those arrangements which support nontrivial explanations are capable of giving rise to explanatory levels in our sense. But to build an explanatory criterion into the levels of organization approach then turns it into a version of our own framework (although one which may be more restricted in terms of which kinds of explanations can be involved). To put things slightly differently: insofar as levels of organization give rise to useful explanations, they may either be identified with, or correlated with, levels of explanation in our sense.

Rival five: mechanistic levels. Mechanistic levels are perhaps the closest existing approach to explanatory levels as we envisage them: they have a constitutive link to explanations of the mechanistic sort (Craver (2007), Andersen (2014a, 2014b)). However, we don't want to restrict the scope of our account only to levels of mechanistic explanation. Mechanistic approaches to explanation in physics (Felline (2022)) remain somewhat underdeveloped; mechanistic approaches get no traction at all in mathematics or metaphysics, or more generally where the relations between levels do not have any compositional character.[3] Our framework for levels

[3] Some chapters of this volume explore levels associated with clearly non-mechanistic explanation: Taylor considers metaphysical explanation, Antos and Colyvan consider mathematical explanations.

of explanation can be, but need not be, combined with an imperialist thesis that all explanations are mechanistic explanations; so explanatory levels are more general than mechanistic levels.

One crucial point which our preferred framework of levels of explanation leaves open (whereas many of the accounts already discussed settle it in some way) is the status of the explanations involved as more or less objective (or mind-independent, or interest-nonrelative). The framework can be combined with robust scientific and metaphysical realisms (as in the chapters by List, Franklin, and Kincaid) or with some variety of pragmatism about explanation (as in the chapters by Chirimuuta, Bechtel, and Hicks). The discussion in Chirimuuta's chapter about the connection between epistemic and ontological levels is relevant here, as is the tripartite distinction between types of levels—descriptive, ontological, explanatory—in List's chapter.

We regard it as a key advantage of explanatory levels over other types of levels frameworks that it leaves the question of explanatory realism open, along with other central metaphysical and epistemological questions. Thinking in terms of levels of explanation does not commit us to any particular approach to explanation; rather, most questions about the status and epistemic role of explanatory levels are outsourced to the broader account of explanation that is combined with the levels framework. As an example, accounts of explanation which make explanation highly context-dependent and localized (and hence not governed by any universal generalizations, as in the classical deductive-nomological (D-N) approach) will typically give rise to highly localized levels structures, as described by Knox, Franklin, and Bechtel. The resulting levels are not characterized neatly by the indiscriminate application of some individual theory or theories: to use another example of Knox's, it makes no sense to talk of 'the sun at the level of statistical mechanics'. Instead, each level is constituted by a patchwork of different theories operating together in a specific modelling setting.

Flexibility about the nature of explanation allows explanatory levels to be a more neutral starting point for levels discussion. This neutrality does, however, come with risks of its own: there may be a temptation to equivocate on the notion of levels, for example in a slide from the modelling of a scientific practice in terms of epistemic levels to the attribution of a levelled structure to reality. Chirimuuta warns of the danger of this slide, while other authors including Kincaid, Knox, and McKenzie seek to place restrictions of various kinds on circumstances in which the inference from epistemic stratification to real explanatory levels is well founded.

The value, or the function, of talk of explanatory levels is a first key theme linking chapters across the volume. Why are levels of explanation useful? We have already discussed some advantages of explanatory levels over other accounts of levels—neutrality on matters of background metaphysics and epistemology, combined with the robust epistemic implications which flow from explanatory levels' constitutive link with explanation. Explanatory levels, on our account, are identified in terms of questions asked and answered, which does not tie them to any specific entity-based or scale-based metaphysics. So explanatory levels may be metaphysically thin, but they are epistemically thick: they have robust links to understanding, prediction, and manipulation.

In our terms, acquiring knowledge of explanations of some phenomenon at multiple different levels *enriches* our explanatory knowledge of the phenomenon. This sense in which multiple explanations at different levels confer richness of explanatory knowledge may be contrasted with the depth of an individual explanation, where (roughly) deeper explanations are those which generalize better (see Hitchcock and Woodward (2003), Weslake (2010)). If we assume a simple link between explanation and understanding—that understanding is a matter of possessing suitable explanatory knowledge—then it immediately follows from what we have said that possessing richer explanatory knowledge gives rise to richer understanding. On more complex accounts of the explanation-understanding link, we may look to account in some different way for the role of explanations at different levels in increasing understanding, but it is plausible that there is a close relationship: better multilevel explanations provide better—in our terms, richer—understanding.

Of course, levels of explanation have their detractors. Some critiques target the coherence of explanatory level frameworks, but more often it is argued that in some sense or other levels of explanation are too narrow a framework in which to fit the complexities of actual scientific inquiry. A clear example of this kind of critique is Kim (2002), who rejects the Oppenheim-Putnam model for requiring a total ordering of levels and instead advocates a more flexible partial ordering, according to which entities can be incommensurable with respect to level, with neither higher level than the other: Kim offers the example of plants and animals. It is also very natural to think that psychology and geology operate at incommensurable explanatory levels. Another style of critique—emphasized by authors as different as Jackson and Pettitt (1990) on programme explanations, Gillett (2016) on machretic explanation, and Potochnik and Yates in their chapters in this volume—is that structural and contextual conditions at different levels may be equally or more crucial to understanding a system's behaviour than external causal agents at the same level as the system.

A common factor in these critiques is that an adequate framework for levels of explanation should not be too rigid; placing metaphysical constraints on what belongs to which level, or on how explanations align with features of some other metaphysical hierarchy, is liable to render a levels concept insufficiently flexible for the full range of applications needed. In light of this sort of consideration, the chapters in this volume which make positive use of levels of explanation tend to employ a flexible concept of level, which invokes partial rather than total ordering (List, Knox, Kincaid) and which allows for multilevel explanation rather than restricting explanations to hold only within individual levels (Franklin, Weslake, Hoffmann-Kolss, Yates).

Adopting a highly flexible conception of levels may avoid various objections, but it risks generating a new kind of over-flexibility problem. If we allow for explanations to range across levels, do we risk losing distinctions between the levels altogether, such that science and metaphysics collectively characterize one single 'wide' level? This is the problem of *too much interaction across levels*: we regard it as an important challenge for detailed spelling-out of theories of explanatory levels. Different approaches will tend to solve this problem in different ways, by identifying some substantive criterion other than mere potential explanatory relevance by which to stratify the different levels. For example, Knox's chapter individuates levels via changes of variable that are physically useful, Strevens's chapter individuates them in terms of probabilistic 'semi-detachment', and the cognitive-science levels of explanation of Marr (1981) (the subject of Chirimuuta's chapter) are individuated in functional terms.[4] Relatedly, some characterizations of explanatory levels have appealed to explanatory proportionality (in the sense of Yablo (1992)) to help individuate levels: even if explanations can hold across levels, the most proportionate explanations might still be between variables at the same level. Wimsatt (1972, 2007) employs a strategy of this kind in terms of 'local' explanations; this strategy is the target of criticism in Potochnik's chapter.

A second major theme of the volume, then, is the opportunities and pitfalls of using models which span different explanatory levels. The existence and usefulness of multilevel explanations refute simplistic approaches to levels according to which explanation always holds within and only with any given level. Typical compositional level approaches tend to sharply distinguish explanation at a given level[5] (typically causation) from explanation across levels (typically grounding or constitution). We take it that one lesson of the literature on levels of explanation is that these types of explanation are not straightforward to disentangle. In different ways, this entanglement is addressed in the chapters by Potochnik, Franklin, Weslake, Hoffmann-Kolss, Yates, and Bechtel.

Suppose that the coherence and applicability of explanatory levels is granted; why think that they are indispensable? One might instead attempt to understand the phenomena we have been aiming to capture simply in terms of the concurrent use of multiple models. Why, it might be asked, do we need to posit anything like levels at all in order to legitimize the use of multiple explanatory models for different theoretical and practical purposes? What we will call the multiple-model approach dispenses with level talk altogether in favour of exclusive talk of the use of multiple models.

[4] The relevant functions here: solving a specific computational problem, connecting specific input to specific outputs, or implementing specific mechanical operations.

[5] Potochnik's chapter refers to this as 'local' explanation.

The difference between explanatory levels and multiple-model approaches ought not to be overstated. Both approaches involve denying that any single explanatory model can tell the whole story about systems of interest; both approaches involve putting together distinct explanatory models in certain ways to give a richer understanding of the target systems; in each approach, we need to take care not to mix levels or models in ways which lead them to break down. Generally, though, levels-based approaches are associated with a stronger commitment to the unity of science, and correspondingly there is some pressure on defenders of these approaches to exhibit in detail the relation between levels. By contrast, multiple-model approaches leave open whether the models are connected in any interesting manner. In particular, levels frameworks typically assume some kind of 'level connector'; a relation between levels, typically transitive, which makes sense of how a higher level can harmlessly coexist with a lower level.

The explanatory level framework in general need not be committed to any specific level connection; indeed, the level connection might vary from case to case. And of course, how to understand this connection is highly controversial amongst level enthusiasts. A range of options for level-connectors have been considered in the literature: supervenience (e.g., List (2019), List's chapter in this volume, Ladyman and Ross (2007), Strevens (2008), Woodward (2021)), reduction (Oppenheim and Putnam (1958), Rosaler (2015, 2019), Crowther in this volume), grounding (Bliss and Trogdon (2021), Bryant (2018)), diachronic emergence (Humphreys (2016), Guay and Sartenaer (2016)), and other more epistemic and pragmatic notions (Dennett (1991), (2009), Chirimuuta in this volume). List's framework for explanatory levels in Chapter 1 is specified in terms of supervenience, and he explicitly argues that this doesn't entail a reductive relation between levels. However, most discussion of levels in physics foregrounds reduction (as in Franklin's and Knox's chapters), whereas contemporary discussions in the metaphysics of science often focus on grounding (as in McKenzie's and Crowther's chapters). What isn't controversial within the explanatory levels approach is that there is *some* link between the levels. This highlights again what we see as one of the underlying functions of levels talk—that it illuminates the different levels of explanation of a phenomenon as aspects of a larger unified explanatory whole, rather than as a collection of dissonant and disconnected fragments of explanation.

Our final theme is the applicability of the levels of explanation framework beyond science to explanations within philosophy—including to those explanations which we give while thinking about explanation itself.[6] Throughout this volume

[6] This kind of turning of theories of explanation on themselves has precedent in the recent literature: Emmerson (forthcoming) compares different metaphysical accounts of explanation in terms of their explanatory depth. Taylor's chapter in this volume also considers the 'explanatory distance' associated with certain explanations in the theory of explanation.

we find candidate explanations being offered for some feature or other of our practice of explaining things at different levels. In several cases, we can think of these candidate explanations as operating at different levels of explanation.

The question of how levels of explanation are possible is closely linked to the broader question of how higher-level explanations are possible at all. Why don't we have to depend on quantum field theory for every explanation we give? Why don't we always use physics for every explanation—why is there *independence* from physics at the higher levels? Fodor (1998) declared himself unable to answer this question;[7] Loewer (2008) makes an attempt at it. We see the chapters in the final section of the volume as collectively contributing to this explanatory project. We face a fraught balancing act: higher levels are usually thought to depend in some way on lower levels (cf. List's discussion of supervenience or Knox's discussion of reduction)—but this dependence mustn't crowd out the independence of the higher level. The conundrum is how to have the right amount of independence (such that there are levels at all) and the right amount of dependence (such that we can make systematic sense of the connections between levels). The fingerprints of this conundrum are found in a variety of related debates: in the reduction-emergence debate, emergence emphasizes higher-level independence, whilst reduction emphasizes the instances of interlevel dependence. Similarly, much of the philosophical interest in effective field theories, and renormalization group methods, as discussed by McKenzie, comes from the possibility that these mathematical methods give us insight into how levels are possible.

To illustrate this theme, the final section, VI, contains three chapters offering different kinds of explanation of how levels of explanation are possible—Bhogal in terms of a pragmatic account of naturalness which makes levels seem more or less inevitable, Strevens in terms of objective probabilistic relations between facts at different scales, and Hicks in terms of features of the underlying Humean laws of nature. These accounts do not directly compete—indeed all three could in principle be combined into a consistent neo-Humean account of explanatory levels and their epistemic role. Accordingly, the three chapters can be seen as complementary explanations, at different levels, of how explanatory levels are possible.

The theme of explaining how levels of explanation are possible connects Section II, which is about the scope and limits of causal modelling at different levels, with Section V, which is about the metaphysical preconditions for level structures. We can factor the preconditions for multilevel explanations into two sorts. We must be able to keep track of the different variables at the different levels and how they are incorporated into a single model; this is the focus of the chapters by Weslake, Hoffmann-Kolss, and Yates. The relevant variables themselves must in addition

[7] 'Well, I admit that I don't know why. I don't even know how to think about why. I expect to figure out why there is anything except physics the day before I figure out why there is anything at all, another (and presumably related) metaphysical conundrum that I find perplexing' (Fodor 1998, p. 161).

actually stand in the right relations for the multilevel explanatory practice to gain traction; this requirement is scrutinized in the chapters by Knox, McKenzie, Emery, Bhogal, Strevens, and Hicks.

Having laid out the main focus and themes of the volume, we turn next to summarizing the individual chapters and identifying their relations to these larger themes. The contributions are grouped into six sections: I) on the foundations of level frameworks, II) on levels of causal explanation, III) on explanatory levels in higher-level sciences, IV) on explanatory levels in physics, V) on levels of explanation in mathematics and in metaphysics, and VI) on the metaphysical conditions which make explanatory levels possible.

Section I of the volume presents and compares some different frameworks for thinking about explanatory levels in the sciences and relating them to other level-based hierarchies. Capturing the relationships between different levels of explanation is a central challenge, but one which is too often answered only schematically or metaphorically. Christian List has in recent work offered a unified formal framework for modelling different types of levels and the relationships between them (List, 2009). In his chapter, List extends this influential framework in several new ways. Typically the relationships between levels have been modelled as supervenience mappings. List's chapter considers and compares multiple different interpretations of the mappings between levels, including supervenience, grounding, and reduction. In particular, he discusses how formal features of grounding such as irreflexity and the potential lack of transitivity fit less easily into his formal framework. List then goes on to explore the conditions under which supervenience entails reducibility. Since Oppenheim and Putnam (1958), level hierarchies have predominantly been 'entity based'; whether compositional or not, the levels are populated by individuals. By contrast, List endorses a 'fact-based' rather than an 'entity-based' conception of levels, and in his chapter he defends this key foundational move. Armed with this precise and formal account of levels, List then goes on to demonstrate the ways in which such precise levels talk can be useful. By being careful about which level a concept belongs to, or at which level a question is asked, new light can be shed on questions ranging from free will to chance and determinism.

Whilst List's chapter considers the technical foundations of explanatory levels, Potochnik's chapter challenges the core motivations for the levels paradigm. A key original motivation for thinking in terms of levels of explanation stemmed from the explanatory anti-reductionist view that explanations in non-fundamental terms are ineliminable, and indeed often provide the best explanation of a given phenomenon. But Potochnik argues that the focus on a hierarchical structure of levels places artificial constraints on our theorizing about scientific explanation and especially about its pragmatic and interest-relative aspects. In particular, the interesting explanations that different sciences generate often don't have any clear relations of metaphysical determination between them; instead different sciences such as neuroscience and biology investigate different causal factors contributing to a particular phenomenon. Sometimes these causal factors will stem from structural, or non-local, factors. Potochnik argues that the plausible thesis of scientific anti-reductionism is not best explicated in terms of levels, since the defining features of different explanations are not readily characterized as defining levels. Instead, Potochnik suggests, there is a great variety of explanations that bear no special relationship to one another but simply feature different influences.

In recent work, Alex Franklin has argued that the value of higher-level explanation needn't rest on a failure of reduction. In his chapter for this volume, Franklin turns to the multiscale modelling practice recently discussed by authors including Robert Batterman, Julia Bursten, and Mark Wilson (Batterman (2013, 2021), Bursten (2018), Wilson (2017)). One theme of this recent work is that the relationships between levels are more complicated than mere averaging of variables, and often unrealistic assumptions are involved, such as assuming the environment to be uniform. The way that models involve different scales is taken by some to imply that the world cannot be neatly divided into levels that are then connected by reduction. This multiscale argument is incompatible with methodological reductionism—which Franklin joins Bursten, Batterman, and Wilson in rejecting—but he asks: is it compatible with a more sophisticated form of reductionism? Franklin answers yes, provided that we are suitably nuanced in our account of reduction. In agreement with proponents of the multiscale argument, Franklin argues that we should pay attention to the complexities of modelling. Doing so will involve local applications of collections of techniques—rather than whole theories—which signals a move away from a Nagelian view of reduction, and a step towards local, and contextual, levels. But there is still room for reductive explanations which do not explain the phenomena of the multiscale model; rather, they explain the effectiveness of that model for its target phenomena. Why we use particular mesoscopic and macroscopic variables should be explicable from the bottom up, if this more nuanced reductionist position is to succeed.

Section II of the volume focuses on the complexities of multilevel causal explanation. The most familiar form of explanation is causal, and in recent years our understanding of causal explanation has been revolutionized by the formal framework of causal modelling and the interventionist interpretation of this framework, associated with Judea Pearl et al. (2000) and James Woodward (2003). Causation is most familiar to us in the higher-level domain of macroscopic objects, stable over time, and interacting with nearby objects; the philosopher's paradigm case is a billiard-ball impact. But causal explanation extends throughout the higher-level sciences, from evolutionary biology to cognitive psychology to astrogeology, and it can hold across levels, as when a cosmic ray causes a cell mutation, or when a human decision causes a click of a mouse. Whilst the recent developments in scientific causal modelling techniques are hugely promising and successful, they are tested to their conceptual limits by application to multilevel phenomena, and an influential objection by Michael Baumgartner casts doubt on the applicability of causal models to such phenomena (Baumgartner (2009)).

In his chapter, which has already seen considerable circulation and discussion in manuscript form, Brad Weslake argues that the popular interventionist account of causation offers a distinctive built-in escape route from Jaegwon Kim's notorious causal exclusion argument (Kim, 1993). Weslake defangs Baumgartner's objection by imposing a condition—the metaphysical possibility of the independent manipulability of the variables in a model—which enables mental and physical variables to be coherently combined into models. This permits a vindication of the possibility of causation of physical effects by mental causes (and vice versa) on both internalist and externalist understandings of mental content. But interventionist accounts of multilevel explanation are not yet out of the woods; Weslake draws attention to a problem raised by Rescorla (2014) which threatens to make high-level interventionist explanations too abundant. In conclusion, Weslake identifies a need to address Rescorla's problem through improving our understanding of what makes a causal model an apt representation of a given causal system.

Aptness conditions are also at the core of Vera Hoffmann-Kolss's searching examination of the prospects for interventionist accounts of multilevel explanation. In her chapter, Hoffmann-Kolss identifies a new class of problem for these accounts, a problem which cannot be addressed by existing interventionist responses to Baumgartner's objection. The trouble is that these accounts tend to trivialize the constraint that models represent all relevant confounding factors. The broader lesson drawn by Hoffmann-Kolss is that in order for the causal modelling framework to succeed in application to multilevel explanation, metaphysically 'thick' notions such as grounding, naturalness, and nomological modality are required—notions which the interventionist has traditionally shunned. Hoffmann-Kolss argues however that some such metaphysical commitments are needed to aptly model multilevel causal structures.

Section II is rounded out by David Yates's chapter, which argues that causal closure of the physical (the idea that all causes are physical causes) can come apart from the causal-explanatory closure of fundamental physics (the idea that all causal explanations reside at the level of fundamental physics). Yates takes vector composition as his core example, arguing that any principle of causal closure that is strong enough to entail the causal-explanatory closure of fundamental physics is falsified by vector composition. Conversely, any weaker principle of causal closure which is compatible with vector composition fails to rule out downward causation. This, Yates argues, makes room for a potential causal role for higher-level properties more generally. The chapter concludes by comparing Yates's approach to Marc Lange's hierarchy of necessity for laws and metalaws, which is built on a network of primitive 'subjunctive facts'.

The third section of the volume turns to higher-level sciences and the explanatory levels that they involve. The chapters of Section III deal with three progressively higher-level sciences—biology, cognitive science, and social science. These sciences are typically taken by realists to stand in some relation of functional realization: cognitive agents are typically realized biologically, while social systems typically involve the interaction of multiple cognitive agents.

The chapter by William Bechtel focuses on the life sciences, and in particular on the notion of a control mechanism as it appears in biological explanations within different domains and at different scales. Bechtel argues that there is no simple over-arching hierarchy of levels of control, but that a level-like notion can nonetheless be recovered in particular cases through consideration of specific self-controlled systems and their components. Bechtel is accordingly cautiously positive about the usefulness of a suitably flexible notion of level in biology: levels of mechanisms and levels of control are important for the explanatory practices of biologists and for philosophers understanding those practices, including both 'top-down' and 'bottom-up' causal claims. He emphasizes however that these levels must remain local and contextual, and that they do not lead to any global stratification of 'biological entities'.

Within the cognitive sciences, levels talk has centred around David Marr's three levels: computation, representation, and implementation (Marr (1982)). Mazviita Chirimuuta's chapter considers how analogies with machines and artefacts (most famously the analogy between the brain and a computer) motivate Marr's three levels, and how this motivation fits into a wider anti-reductionist view of explanation in brain and behavioural sciences. Chirimuuta raises some challenges for the analogy between designed artifacts like a radio and evolved systems like the central nervous system. Chirimuuta emphasizes the explanatory importance of these levels: oftentimes a problem is intractable if we start with all the details, but just as analogies might lead us astray by framing problems/phenomena in an overly simplistic way, the levels structure might be a heuristic that is helpful only to a limited extent. Then we run the risk of ending up attributing levels not just to our representations, but to reality itself—thus projecting methodological levels into metaphysical levels.

The final chapter of Section III, by Harold Kincaid, concerns explanatory levels in the social and behavioural sciences. The complexity and messiness of theorizing in social science may suggest that there will be no neat application of any explanatory levels framework; but Kincaid argues that levels can nonetheless be distinguished and used successfully by social scientists. Kincaid emphasizes taking a naturalist 'science-first' approach, one which emphasizes how explanatory levels are a contextual matter. Kincaid argues that macrosociological entities are not epiphenomena by emphasizing their causal roles in a range of different case studies, including some which he argues are benign examples of downwards causation. The result is again a flexible approach to explanatory levels which is motivated by the need to make sense of the full range of scientific practice rather than by any metaphysical precepts.

Section IV turns to explanatory levels in physics. As Eleanor Knox emphasizes in her chapter, connections between levels are often relatively well understood in physics in comparison to other sciences. The tools of mathematics allow us to wield precise control over transitions between physical explanatory levels, as in the important case of renormalization techniques in quantum field theory and condensed matter physics. Oftentimes we know how to construct one physical explanatory level from another in a way that would seem fanciful in the mind/brain case—even to the point where we can claim full-scale reduction (Robertson (2022)).[8] Still, explanatory levels in physics otherwise have much in common with explanatory levels in other sciences. Since hardly any of physics is fundamental physics, as Knox emphasizes, most theorizing in physics takes place at an effective level; many of the same issues then arise for explanatory levels in physics as for levels in biology and other higher-level sciences. This helps cast light on how apparently conflicting physical theories at different levels can be reconciled: McKenzie uses the example of how symmetries which hold at one physical level may be broken at another.

We have already raised a concern about whether explanatory levels might be too easily achieved. In her chapter, Eleanor Knox asks: what makes for a thicker notion of levels? (Or, as she puts it: 'levels worth having'.) Knox emphasizes the contrast between mere levels of description, and the more interesting levels that we come across in physics. In typical cases of levels in physics, the level connection is well understood, but might crucially involve changes of variable to reveal new dependencies. Here her position contrasts with the related approach of Franklin, according to whom explanatory levels in physics are cheap and plenitudinous. Emergence has often been given a two-part treatment: in terms of robustness/autonomy and then in terms of novelty. Much of the discussion in the philosophy of physics has focused on the former—and Knox suggests that to understand the contrast between thin and thick levels, we should turn our attention to novelty. Building on her previous work, Knox proposes that in contrast to the subjective/psychological view of Butterfield (who analyses 'novelty' as 'surprise'), we should understand novelty as novel explanatory value.

Discussions of levels often presuppose the level of fundamental physics as given unproblematically; the truth is far more complicated. Current physics neither offers us a plausible candidate fundamental theory, nor any clear guidance as to what such a theory would look like. Moreover, metaphysics is sometimes prone to usurp the claims of physics to fundamentality: as already discussed, metaphysically

[8] The improved prospects for reduction in physics have led a number of philosophers of physics to the conclusion that reduction and emergence can be compatible (Wilson (2010), Butterfield (2011b), Crowther (2015)). However, to accommodate this compatibility, the relevant notion of emergence will tend to be a weaker one (see, e.g., the distinction between strong and weak emergence in Chalmers (2006)). The explanatory levels paradigm thus ties in with sophisticated recent discussions of emergence in physics such as Franklin and Knox (2018). Still, if the levels involved are too cheap or abundant, then they might not be strong enough to support even a weak notion of emergence.

orientated philosophers of science have explored the prospect of giving metaphysical reductions of physical concepts like 'length' and 'mass'. This raises the question as to whether the levels of explanation of metaphysics and physics can be brought into line. Do the levels of fundamentality acknowledged in recent metaphysics line up with the levels of explanation given by different physical theories? The two chapters by Karen Crowther and Kerry McKenzie tackle this problem head-on—and give opposing answers.

Karen Crowther sees explanatory levels in metaphysics and physics as more in harmony than McKenzie does. Crowther aims to bring metaphysics and physics closer in line, by employing a new definition of 'relative fundamentality' according to which theoretical descriptions with broader scope can count as more fundamental than theoretical descriptions which can be derived from them in specific domains. This new definition aims to shift the emphasis in the metaphysics of physics from questions about the mereological structure of levels to questions about the relations of reduction and constructability that hold between different theoretical descriptions. Crowther argues that reduction relations between theories operating at different scales are often thought to have ontological import, in particular she holds that derivability can be understood as natural, or, indeed, ontological dependence. But the radical part of her argument is that we should also think of reduction relations between old and new theories as similarly having ontological import. This is a version of an idea which has gained traction in recent years, sometimes under the guise of 'effective realism' (Williams (2019), Egg (2021), Saatsi (2022), Robertson and Wilson (forthcoming)).

McKenzie also focuses on the nature of interlevel relationships in theoretical physics, with a sustained critique of applications of the metaphysics of grounding to the topic. McKenzie identifies a tension between the approximation inherent in interlevel relations in physics and the exactness presupposed by standard metaphysical accounts of interlevel grounding. This, she argues, undermines any attempt to use the levelled structure of scientific explanations to support a metaphysically heavy-duty account of levels of explanation in grounding terms. The key problem is that the approximations involved in understanding inter-theoretic relations end up being interest relative, and hence not completely objective in the way that grounding is meant to be. McKenzie's lead example is proton decay—but it can be extended to applications of the effective field theory framework more generally. The tension McKenzie identifies is exacerbated since effective field theories are sometimes regarded as helping to explain *why* we have different levels in the first place; this latter type of question is revisited in Section VI.

Section V broadens discussions of explanatory levels to new questions and domains within mathematics and metaphysics. As McKenzie's chapter already highlights, there is a kind of motivation present here that is at least partly separate from the hierarchy of different sciences. The centrality of fundamentality in metaphysics and of axiomaticity in mathematics build the idea of one thing being at a more basic level than another right into the self-conceptions of the relevant fields. The chapters in this section contrast the multiplicity of kinds of explanations found in mathematics and in metaphysics with the classical model of explanatory levels.

In their chapter, Carolin Antos and Mark Colyvan explore cases where different styles of mathematically explanatory proof exist for a given result. Taking Fermat's Little Theorem and Descriptive Set Theory as examples, Antos and Colyvan show that different styles of proof of a single result can exhibit different balances of theoretical virtues, with mathematicians' intuitions about explanatory power favouring more general styles of proof in some cases and more local styles of proof in others. This verdict reinforces the volume's theme that explanatory levels can be useful even when applied piecemeal and in the absence of any global ordering.

The next chapter turns to explanatory levels within metaphysics. Elanor Taylor raises a worrying objection to recent accounts of dispositions as grounded in lower-level metaphysical facts. Taylor argues that these explanations look very similar to 'dormitive-virtue' style explanations, the archetypal example of unsuccessful explanation. But what exactly is wrong with dormitive-virtue explanations? Taylor teases apart the good from the bad in these explanations, and argues that the dispositionalist's explanations can be partially vindicated by appeal to a 'backing' model of explanation—a model which appeals to different dependence relations in different explanatory contexts within metaphysics.

Completing Section V is Nina Emery's chapter, which criticizes existing accounts of higher-level chances as unable to do justice to their explanatory power. Emery then offers a new account of the nature of higher-level chances with better prospects of vindicating the explanatory role which leads us to postulate those chances in the first place. The proposal does however present a challenge for Humean accounts of the role of chance in a deterministic world: the specific way in which Humeans take chances to depend on the mosaic of fundamental facts compromises the chances' ability to play the explanatory role they are assigned.

The volume thus far has discussed what explanatory levels might look like, and what use they might be in a variety of different settings. Even the most sceptical about levels tend to agree that the different sciences can investigate the world independently of one another, to some extent at least. Yet it seems like a possibility that the world could have been different in this respect. It apparently could have been the case that geological questions couldn't after all be answered without in-depth reference to quantum physics, or that no useful social science generalization could be formulated without delving into the details of individual psychology. So: why are there distinct levels of explanation in the first place?

The final section, VI, takes a step back and asks about the pre-conditions for levels of explanation. This returns us to the volume's third general theme: levels of explanation for levels of explanation. A foundational challenge, yet one which is rarely confronted head on, is to explain why there are any high-level regularities at all. As we discussed above, Jerry Fodor (1974, 1997) posed this question, but there is no consensus about what the answer should be. In the first chapter of this section, Harjit Bhogal offers a distinctive new solution to Fodor's challenge which draws in part on our pragmatic interests and projects as scientific reasoners. Bhogal points out that Fodor's challenge presupposes an identification of certain high-level properties as natural properties, and then shows that the existence of higher-level regularities and of the corresponding explanations falls straight out of some plausible contemporary accounts of natural properties—in particular, out of Bhogal's own proposal and out of an influential recent account by Barry Loewer (2020). Bhogal concludes by contrasting two complementary levels of explanation of high-level regularities: a 'bottom-up' scientific strategy in terms of the specific features of the low-level realizers, and a 'top-down' metaphysical strategy in terms of the nature of properties.

In the following chapter, Michael Strevens offers a general version of Bhogal's bottom-up strategy, a version which highlights the core role of stochastic independence between variables at different levels in underwriting high-level explanations. Strevens outlines a schematic theory of the possibility of high-level explanation in terms of *semi-detachment* of variables at different levels, a theory which is compatible with the various ways in which these variables are causally entangled. In emphasizing independence between variables at different levels, Strevens's discussion dovetails with Knox's discussion of levels in physics. But why is semi-detachment so common? Strevens next turns to the possibility of providing a top-down explanation of semi-detachment itself, in terms of the propensity of the details of lower-level systems to cancel or balance one another out. He argues that this propensity to cancel out explains why high-level explanations in the sciences are so widespread.

A full account of the explanatory power of higher-level sciences not only needs to address the properties and regularities as Bhogal does, and to demonstrate their relationship to different levels as Strevens does; it also needs to encompass the explanatory role of the related higher-level laws—and, as we saw in Emery's chapter in the previous section, higher-level chances. Although best-system accounts of laws in the tradition of Mill, Ramsey, and Lewis are deservedly popular, they tend to face difficulties in accounting for laws other than those which hold at the fundamental level, and in making sense of the explanatory role of chance. Higher-level laws are a central concern of explanatory levels; for those who hold that explanation hinges on laws, including proponents of the D-N model and its descendants,[9] without higher-level laws there would be no distinctive higher-level explanations. The law-like status of regularities in the higher-level sciences is central to the Fodor-Oppenheim-Putnam debate and the genesis of explanatory levels; one of

[9] See Woodward and Ross (2021) for a survey.

Fodor's key objections to the reducibility of the higher level stems from whether the lower level can explain/account for lawhood at the higher level.[10]

This theme is taken up in the final chapter of the volume, in which Michael Townsen Hicks offers a new and improved version of the popular 'better best system' analysis of laws at higher levels. Craig Callender and Jonathan Cohen (2009), and separately Markus Schrenk (2014), have attempted to characterize special-science laws by modifying the best-system approach so as to apply at multiple levels. Hicks improves on these proposals by accommodating systematic interlevel relationships without undermining the explanatory power of higher-level laws: the key is to assess a candidate law's informativeness in terms of facts at multiple different levels of interest. In this way, Hicks offers a *big* better best system that he terms the 'democratic view' of laws. This democratic view treads a fine line between those that hold that the higher-level laws are independent (anarchists) and those that hold that the more fundamental laws are responsible not only for the regularities of higher-level laws but also for their lawhood status (imperialists). Imperialists require too strong a connection between higher and lower levels, one which leaves features like the counterfactual robustness of higher-level laws unexplained: they have no account of how the economic law of supply and demand would still have held even with a different periodic table of the elements. On the other hand, anarchists have no good account of how we achieve detailed understanding of interlevel relations in many cases; the need to account for this epistemic success was particularly emphasized throughout Section V, as well as in Franklin's discussion of reductive explanations.

Taken together, the chapters of Section VI give us insight into what underlying aspects of our world allow the different levels of explanation to be possible in the first place. They thereby offer explanations at various different levels for the multifaceted phenomenon of levels of explanation.

References

Andersen, Holly (2014a). "A Field Guide to Mechanisms: Part I," *Philosophy Compass* 4: 274–283.

Andersen, Holly (2014b). "A Field Guide to Mechanisms: Part II," *Philosophy Compass* 4: 283–297.

Batterman, Robert W. (2013). "The Tyranny of Scales." In *The Oxford Handbook of Philosophy of Physics*, ed. Robert W. Batterman. Oxford University Press, pp. 256–286.

Batterman, Robert W. (2020). "Multiscale Modeling: Explanation and Emergence." In *Methodological Prospects for Scientific Research*, ed. Wenceslao J. Gonzalez. Springer, pp. 53–65.

Batterman, Robert W. (2021). *A Middle Way: A Non-Fundamental Approach to Many-Body Physics*. Oxford: Oxford University Press.

Baumgartner, Michael (2009). "Interventionist Causal Exclusion and Non-Reductive Physicalism," *International Studies in the Philosophy of Science* 23(2): 161–178.

[10] See Sober (1999) for a response.

Bechtel, William (1994). "Levels of Description and Explanation in Cognitive Science," *Minds and Machines* 4(1): 1–25.

Bechtel, William (2008). *Mental Mechanisms: Philosophical Perspectives on Cognitive Neuroscience*. London: Routledge.

Bennett, Karen (2017). *Making Things Up*. New York: Oxford University Press.

Bliss, Ricki, and Kelly Trogdon (Winter 2021 Edition). "Metaphysical Grounding." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University, available at https://plato.stanford.edu/archives/win2021/entries/grounding/.

Brooks, Daniel Stephen, James DiFrisco, and William C. Wimsatt (eds.) (2021). *Levels of Organization in the Biological Sciences*. Cambridge: MIT Press.

Bryant, A. (2018). "Naturalizing Grounding: How Theories of Ground Can Engage Science," *Philosophy Compass* 13(5): e12489.

Bursten, Julia R. (2018). "Conceptual Strategies and Inter-theory Relations: The Case of Nanoscale Cracks," Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 62: 158–165.

Butterfield, Jeremy (2011a). "Emergence, Reduction and Supervenience: A Varied Landscape," *Foundations of Physics* 41(6): 920–959.

Butterfield, Jeremy (2011b). "Less Is Different: Emergence and Reduction Reconciled," *Foundations of Physics* 41(6): 1065–1135.

Butterfield, Jeremy (2014). "Reduction, Emergence, and Renormalization," *Journal of Philosophy* 111(1): 5–49.

Callender, Craig, and Jonathan Cohen (2009). "A Better Best System Account of Lawhood," *Philosophical Studies* 145(1): 1–34.

Cartwright, Nancy (1999). *The Dappled World: A Study of the Boundaries of Science*. Cambridge: Cambridge University Press.

Castellani, Elena (2000). "Reductionism, Emergence, and Effective Field Theories," *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 33(2): 251–267.

Chalmers, David J. (2006). "Strong and Weak Emergence." In *The Re-Emergence of Emergence: The Emergentist Hypothesis From Science to Religion*, eds. P. Davies and P. Clayton. Oxford University Press, pp. 244–254.

Craver, Carl. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*, Oxford: Oxford University Press.

Craver, Carl. F. (2015). "Levels." In *Open MIND*, eds. T. Metzinger and J. M. Windt. Frankfurt am Main, Germany: MIND Group, pp. 1–26.

Crowther, Karen (2015). "Decoupling Emergence and Reduction in Physics," *European Journal for Philosophy of Science* 5(3): 419–445.

Dennett, Daniel C. (1991). "Real Patterns," *Journal of Philosophy* 88(1): 27–51.

Dennett, Daniel C. (2009). "Intentional Systems Theory." In *The Oxford Handbook of Philosophy of Mind*, eds. Ansgar Beckermann and Brian P. McLaughlin. Oxford University Press, pp. 339–350.

Dupré, John (1993). *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge: Harvard University Press.

Egg, Matthias (2021). "Quantum Ontology Without Speculation," *European Journal for Philosophy of Science* 11(32).

Emmerson, Nicholas (2022). "Plumbing Metaphysical Explanatory Depth," Philosophical Studies: 1–22. https://doi.org/10.1007/s11098-022-01886-3

Eronen, Markus I. (2015). "Levels of Organization: A Deflationary Account," *Biology and Philosophy* 30(1): 39–58.

Felline, Laura (2021). "Mechanistic Explanation in Physics". In *The Routledge Companion to the Philosophy of Physics*, eds. Eleanor Knox and Alastair Wilson. Routledge, pp. 476–486.

Fine, Kit (2012). "Guide to Ground." In *Metaphysical Grounding*, eds. Fabrice Correia and Benjamin Schneider. Cambridge University Press, pp. 37–80.

Fodor, Jerry (1974). "Special Sciences (or: The Disunity of Science as a Working Hypothesis)," *Synthese* 28(2): 97–115.

Fodor, Jerry (1997). “Special Sciences: Still Autonomous after all These Years,” *Philosophical Perspectives* 11: 149–63.

Franklin, Alexander (2018). “Whence the Effectiveness of Effective Field Theories?,” *British Journal for the Philosophy of Science* 71(4): 1235–1259.

Franklin, Alexander, and Eleanor Knox (2018). “Emergence Without Limits: The Case of Phonons,” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 64: 68–78.

Halpern, Joshua Y., and Judea Pearl (2005). “Causes and Explanations: A Structural-Model Approach. Part I: Causes,” *British Journal for the Philosophy of Science* 56(4): 843–887.

Guay, Alexandre, and Olivier Sartenaer (2016). “A New Look at Emergence. Or When After Is Different,” *European Journal for Philosophy of Science* 6(2): 297–322.

Hempel, Carl Gustav (1965). *Aspects of Scientific Explanation* (Vol. 1). New York: Free Press.

Hitchcock, Christopher, and James Woodward (2003). “Explanatory Generalizations, Part II: Plumbing Explanatory Depth,” *Noûs* 37(2): 181–199.

Humphreys, Paul (2016). *Emergence: A Philosophical Account*. New York: Oxford University Press.

Kim, Jaegwon (1993). “The Nonreductivist’s Troubles with Mental Causation.” In *Supervenience and Mind: Selected Philosophical Essays*, ed. Jaegwon Kim. Cambridge: Cambridge University Press, pp. 336–357.

Kim, Jaegwon (2002). “The Layered Model,”

Knox, Eleanor (2016). “Abstraction and its Limits: Finding Space for Novel Explanation,” *Noûs* 50(1): 41–6.

List, Christian (2019). “Levels: Descriptive, Explanatory, and Ontological,” *Noûs* 53(4): 852–883.

Lewis, David (1988a). “Statements Partly About Observation,” *Philosophical Papers* 17: 1–31.

Lewis, David (1988b). “Relevant Implication,” *Theoria* 54(3): 161–174.

Loewer, Barry (2001). “From Physics to Physicalism”. In *Physicalism and its Discontents*, eds. Carl Gillett and Barry Loewer. Cambridge: Cambridge University Press, pp. 37–56.

Loewer, Barry (2008). “Why There Is Anything Other Than Physics.” In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, eds. J. Hohwy and J. Kallestrup. Oxford: Oxford University Press, pp. 149–163.

Loewer, Barry (2020). “The Package Deal Account of Laws and Properties,” *Synthese* 199(1–2): 1065–1089.

Nagel, Ernest (1935). “The Logic of Reduction in the Sciences,” *Erkenntnis* 5: 46–52.

Nagel, Ernest (1949). “The Meaning of Reduction in the Natural Sciences.” In *Science and Civilization*, ed. R. C. Stouffer. Madison: University of Wisconsin Press, pp. 99–135.

Nagel, Ernest (1961). *The Structure of Science. Problems in the Logic of Explanation*. New York: Harcourt, Brace & World, Inc.

Nagel, Ernest (1970). “Issues in the Logic of Reductive Explanations.” In *Mind, Science, and History*, eds. H. E. Kiefer and K. M. Munitz. Albany: SUNY Press, pp. 117–137.

Nickles, Thomas (1973). “Two Concepts of Intertheoretic Reduction,” *Journal of Philosophy* 70(7): 181–201.

Marr, David (1982). *Vision*. W. H. Freeman: San Francisco.

Mundy, Brett (1987). “The Metaphysics of Quantity,” *Philosophical Studies* 51(1): 29–54.

Oppenheim, Paul, and Hilary Putnam (1958). “Unity of Science as a Working Hypothesis,” *Minnesota Studies in the Philosophy of Science* 2: 3–36.

Owens, David (1989). “Levels of Explanation,” *Mind* 98(389): 59–79.

Pearl, Judea (2000). *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University.

Perry, Zee (forthcoming). “On Mereology and Metricality,” to appear in *Philosophers’ Imprint*.

Robertson, Katie (2022). “In Search of the Holy Grail: How to Reduce the Second Law of Thermodynamics,” *British Journal for the Philosophy of Science* 73(4): 987–1020.

Robertson, Katie, and Alastair Wilson (forthcoming). “Theoretical Relicts: Progress, Reduction and Autonomy,” *British Journal for Philosophy of Science* 77, available at https://doi.org/10.1086/724445.

Rosaler, Joshua (2015). "Local Reduction in Physics," *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 50: 54–69.

Rosaler, Joshua (2019). "Reduction As an A Posteriori Relation," *British Journal for the Philosophy of Science* 70(1): 269–299.

Saatsi, Juha (2022). "(In)effective Realism?," *European Journal for Philosophy of Science* 12: 30.

Schaffner, Kenneth (1967). "Approaches to Reduction," *Philosophy of Science* 34: 137–147.

Schrenk, Markus (2014). "Better Best Systems and the Issue of CP-Laws," *Erkenntnis* 79(10): 1787–1799.

Sober, Elliott (1999). "The Multiple Realizability Argument Against Reductionism," *Philosophy of Science* 66(4): 542–564.

Strevens, Michael (2008). *Depth: An Account of Scientific Explanation*. Cambridge: Harvard University Press.

Weslake, Brad (2010). "Explanatory Depth," *Philosophy of Science* 77(2): 273–294.

Wallace, David (2019). "Naturalness and Emergence," *The Monist* 102(4): 499–524.

Williams, Porter. (2019). "Scientific Realism Made Effective," *British Journal for the Philosophy of Science* 70(1): 209–237.

Wilson, Alastair (2018). "Metaphysical Causation," *Noûs* 52(4): 723–751.

Wilson, Jessica (2010). "Non-Reductive Physicalism and Degrees of Freedom," *British Journal for the Philosophy of Science* 61(2): 279–311.

Wilson, Mark (2010). "Mixed-Level Explanation," *Philosophy of Science* 77(5): 933–946.

Wilson, Mark (2017). *Physics Avoidance: Essays in Conceptual Strategy*. Oxford University Press.

Wimsatt, William C. (1972). "Complexity and Organization." In *PSA: Proceedings of the Biennial meeting of the Philosophy of Science Association*, eds. K. F. Schaffner and R. S. Cohen. Dordrecht Reidel, pp. 67–86.

Wimsatt, William C. (1976). "Reductionism, Levels of Organization, and the Mind-Body Problem." In *Consciousness and the Brain*, ed. Gordon G. Globus. Plenum Press, pp. 205–267.

Wimsatt, William C. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge MA: Harvard University Press.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press.

Woodward, James (2021). "Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance," *Synthese* 198: 237–265.

Woodward, James, and Lauren Ross (Summer 2021 Edition). "Scientific Explanation." In *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. Stanford, CA: Stanford University. Available at: https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/

Yablo, Stephen (1992a). "Mental Causation," *Philosophical Review* 101: 245–280.

Yablo, Stephen (2014). *Aboutness*. Princeton: Princeton University Press.

PART I

FOUNDATIONS OF EXPLANATORY LEVELS

1

Levels of Description and Levels of Reality

A General Framework

Christian List

1 Introduction

An important feature of science is its organization into different domains of enquiry. In different such domains, we focus on different phenomena and use different concepts and categories to describe and explain those phenomena. Some areas of science focus on larger-scale phenomena—think of astronomy, ecology, or macroeconomics, for instance—while others focus on smaller-scale phenomena, such as particle physics, molecular biology, or microeconomics. We then say that these areas of science operate at different "levels of description" or different "levels of explanation." Some operate at what we call a "micro-level," while others operate at a "macro-level."

But what are "levels"? Although talk of "levels," such as "levels of description," "levels of explanation," or even "levels of reality," is very common in both science and philosophy,[1] and there are many debates on what the right level of explanation is for certain phenomena, such as for social, psychological, or biological ones, this talk of levels is sometimes criticized for being too metaphorical and imprecise. As Jaegwon Kim writes, "talk of levels may turn out to be only a figure of speech, a harmless but suggestive metaphor."[2]

We may have an intuitive grasp of what it means to say that macroeconomics operates at a higher level than microeconomics, or that systems biology operates at a higher level than cell biology, but despite the wealth of relevant scholarly work, there is still no consensus among scientists and philosophers on how to make those claims precise. Further, there is no consensus on how higher-level phenomena or explanations are related to lower-level ones, and whether the former are somehow "reducible" to the latter, at least in principle. Finally, there is no consensus on

[1] See, among many others, Oppenheim and Putnam (1958), Bunge (1960, 1977), Fodor (1974), Owens (1989), Beckermann, Flohr, and Kim (1992), Dupré (1993), Bechtel (1994), Wimsatt (1994), Kim (1998, 2002), Schaffer (2003), Craver (2007), Floridi (2008), Rueger and McGivern (2010), Potochnik and McGill (2012), Ellis, Noble, and O'Connor (2012), Knox (2016), and Eronen and Brooks (2018).

[2] See Kim (2002, p. 3).

Christian List, *Levels of Description and Levels of Reality* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Christian List 2024. DOI: 10.1093/oso/9780192862945.003.0002

whether "levels" should be understood only in epistemic terms, as a feature of how we think about the world, or also in ontic terms, as a feature of reality itself.

Building on the rich body of work in this area, the aim of this chapter is to present a general framework for representing levels and inter-level relations. The framework is intended to capture both epistemic and ontological notions of levels and to clarify the sense in which levels of explanation might or might not be related to a levelled ontology. Moreover, the framework is intended to allow us to study and compare different kinds of inter-level relations, especially supervenience and reduction but also grounding and mereological constitution. This, in turn, will enable us to explore questions such as whether supervenience implies explanatory reducibility and whether there can be irreducible higher-level explanations or even "emergent" higher-level properties.[3]

I will first review several salient uses of the idea of levels, beginning with levels in the epistemic sense (Section 2), followed by levels in the ontic sense (Section 3). I will then show how to accommodate these different notions in a unified framework (Section 4). Next, I will use the framework to address some key questions about the relationship between epistemic and ontic notions of levels (Section 5). Finally, I will briefly mention some other theoretical payoffs and illustrative applications, namely to the free-will debate, the distinction between determinism and indeterminism, indexicality and consciousness, and the relationship between positive and normative facts (Section 6).

2 Levels in the epistemic sense

I will begin with an account of levels in the epistemic sense, i.e., levels of description or levels of explanation, which seems to be the least controversial sense, and only subsequently turn to levels in the ontic sense, i.e., levels of reality, which seems to be more controversial.

As already noted, we use different concepts and categories when we describe and explain the phenomena in different domains. For example, fundamental physics speaks of particles, fields, and forces; biology speaks of cells, organisms, and ecosystems; psychology speaks of mental states, intentionality, and cognition; and the social sciences speak of institutions, norms, and conventions.

[3] This chapter builds on some material in, and is a sequel to, List (2019a), where the proposed unified framework was previously introduced. The present exposition is new and updated in some respects. Among other things, I explicitly consider a greater variety of inter-level relations than I did in that earlier work. Most importantly, however, the present contribution stands on the shoulders of other works. The works cited in Note 1, for instance, have done much to systematize our understanding of different notions of levels. Also note that the literature contains earlier taxonomies or formal analyses of levels, such as in Craver (2007, Ch. 5) and Bunge (1960, 1977). I thank an anonymous reviewer for drawing my attention to some of those cited works.

We have very good explanatory reasons for following this differentiated explanatory practice. Different *explananda*—different phenomena to be explained—require different explanatory concepts and categories, which enable us to recognize different patterns and regularities in the world. It should be evident, for instance, that explaining the movement of the planets in physics or photosynthesis in biology requires very different conceptual resources than explaining inflation in economics or voting behaviour in politics.

"Operating at a different level of description or different level of explanation" simply means describing and explaining the world through the lens of a different system of concepts and categories. A *level of description* or a *level of explanation* can thus be informally defined as a particular system of concepts and categories through which one might describe and/or explain the world. For instance, the fundamental physical level is defined by the concepts and categories of fundamental physics, such as particles, fields, and forces, and the psychological level is defined by a system of psychological concepts and categories, such as beliefs, desires, and other mental states and processes.

Now, there are at least two ways in which such an epistemic understanding of levels can be made more precise.

- *The coarse-graining understanding*: this is based on the idea that each level of description corresponds to a particular way of partitioning some underlying set of possibilities into equivalence classes.
- *The linguistic understanding*: this is based on the idea that each level corresponds to a particular level-specific descriptive or explanatory language.

Let me explain these in turn.

2.1 Levels as equivalence relations

On a "coarse-graining understanding," different levels correspond to different ways of partitioning some underlying set of possibilities—for instance, the set of all possible worlds or the set of all possible states of the world—into equivalence classes. Formally, each level thus corresponds to a particular equivalence relation on that set. An *equivalence relation* on a given set specifies, for any two of its elements, whether they count as equivalent according to the standard encoded by that relation.[4]

Understanding levels as equivalence relations captures the idea that the concepts and categories available at different levels allow us to draw different distinctions in

[4] Formally, an equivalence relation is a reflexive, symmetrical, and transitive binary relation on the given set.

the world and force us to ignore others. Specifically, at each level, the level-specific concepts and categories allow us to distinguish between possibilities that lie in different equivalence classes but not between possibilities that lie within the same equivalence class.

David Lewis already introduced this way of representing levels of description, albeit without using the terminology of "levels." Specifically, he introduced the notion of a "subject matter," which is essentially the same as a level in the present sense.[5] A *subject matter*, for Lewis, picks out a part—or perhaps better, an aspect—of the world, namely the one that has to do with that subject matter. Formally, Lewis takes each subject matter to be representable by an equivalence relation on the set of possible worlds. Physics, biology, and psychology, for instance, are all subject matters under this definition; they each partition possibilities differently, thereby focusing on different distinctions. Two worlds are indistinguishable with respect to physics, or biology, or psychology if and only if they coincide with respect to all physical, all biological, or all psychological properties, respectively.

Lewis also introduces the notion of "inclusion of subject matters."[6] One subject matter is said to *include* another if the equivalence relation representing the former is at least as fine-grained as the equivalence relation representing the latter, i.e., any two possibilities that are distinguished by the latter equivalence relation are also distinguished by the former. So, whenever one subject matter includes another, any distinction that can be drawn in terms of the latter (the included subject matter) can also be drawn in terms of the former (the including one).

Similarly, in some parts of economics and psychology, an agent's *awareness* is sometimes characterized in terms of the distinctions that this agent is able to draw and formally defined as an equivalence relation on some underlying set of possibilities.[7] The agent is said to be *aware* of some feature of the world if and only if he or she can distinguish worlds with that feature from worlds without it. Greater awareness corresponds to a more fine-grained partition, and lesser awareness to a more coarse-grained one. Awareness growth would involve fine-graining. Levels of awareness can again be related to each other by an inclusion relation, defined as in Lewis's account of subject matters.

Inclusion as defined by Lewis and applicable also to awareness is our first example of an inter-level relation. We can say that one level counts as "higher" than another if the equivalence relation representing the former is strictly more coarse-grained than the equivalence relation representing the latter. The inclusion relation ("at least as fine-grained as") yields a partial ordering over all Lewisian subject matters (or levels as equivalence relations), defined for some underlying set of possible worlds.

[5] See Lewis (1988).

[6] Ibid.

[7] See, e.g., Modica and Rustichini (1999) and Dietrich (2018).

At this point, we can already make the first substantive observation: levels in the epistemic sense need not be totally ordered. That is, we shouldn't think of there being a linear hierarchy of levels. Rather, there may be only a partial ordering. Some levels may be comparable in terms of the "higher than" relation, others not. For instance, the levels of biology and geology may each be higher than the level of physics, but neither of them may be higher than the other.

Some people may therefore prefer to speak of "scales," "domains," "conceptual schemes," or indeed Lewisian "subject matters" instead of "levels," but since talk of "levels" is ubiquitous, I propose to retain this terminology, despite the lack of a linear hierarchy.[8]

2.2 Levels as descriptive or explanatory languages

Let me turn to the second way in which levels in the epistemic sense can be made more precise. Here, different levels correspond to different level-specific languages for describing and/or explaining the world. To provide a simple formalization of this, let me define a descriptive or assertoric *language*, **L**, as the set of all (declarative) sentences that can be expressed in it (this includes all sentences that assert propositional content but excludes, for instance, questions and commands), where this language is endowed with (i) some *logical operations*, at a minimum a negation operator, such that, for each sentence in **L**, its negation is also in **L**, and (ii) a well-behaved *notion of consistency*, which partitions the set of all subsets of **L** into those that are consistent and those that are inconsistent. (The latter, in turn, also allows us to define a notion of *logical entailment*.[9]) The simplest examples of such languages come from standard propositional logic, but we could also consider more expressive languages, which may include not only predicates, but also modal operators (such as "necessarily" and "possibly") and/or non-material conditionals (such as "if X were the case, then Y would be the case"). The idea is that such a language **L** can be used to express descriptions or explanations at the given level.

If different levels correspond to different languages in the present sense, we can now also introduce one salient kind of inter-level relation for such levels, namely the reduction relation. One language **L** is *reducible* to another language **L**′ if there

[8] These alternative terms appear, e.g., in Wilson (2010), Potochnik and McGill (2012), Kim (2002), Davidson (1973), and Lewis (1988).

[9] For this notion of a language, see Dietrich (2007). To count as *well-behaved*, the notion of consistency must satisfy the following conditions: first, any sentence-negation pair is inconsistent; second, any subset of any consistent set is still consistent; third, the empty set is consistent and every consistent set has a consistent superset containing a member of each sentence-negation pair within the language. We can then further say that a set of sentences *logically entails* another sentence if the set together with the negation of the sentence is inconsistent.

exists a *translation function f* from **L** to **L**′ which assigns to each sentence ϕ in **L** an "equivalent" sentence $\phi' = f(\phi)$ in **L**′, where logical properties (such as consistency, inconsistency, and negation) are preserved under translation.

For example, if we had a function that assigns to each sentence expressible in the language of chemistry a content-wise equivalent sentence in the language of physics, then we would have achieved a reduction of chemistry to physics. It is a non-trivial question, however, whether, and under what conditions, such reductions exist, and I will say more about it in Section 5. For the moment, I want to note that even the question of whether chemical descriptions are reducible to physical ones—a familiar example of purported reducibility—is controversial.[10] Again, different levels in the present sense are partially, but not completely, ordered by the given inter-level relation.

We may also ask how the two epistemic notions of levels I have introduced—levels as equivalence relations and levels as descriptive or explanatory languages—are related to one another, and similarly how their respective inter-levels relations are related. As should become clear, the framework to be presented will offer some formal tools for addressing those questions.

3 Levels in the ontic sense

Let me move on to the ontic understanding of levels. Here the idea is that levels are not merely a feature of our way of thinking about the world and describing it, but a feature of reality itself. According to a levelled ontology, the world is somehow stratified into levels. In line with such a picture, philosophers often invoke notions such as "the fundamental level of reality." And if one speaks of the fundamental level of reality, then presumably it also makes sense to speak of other, higher levels. As Jonathan Schaffer, for instance, observes: "[t]alk about 'the fundamental level of reality' pervades contemporary metaphysics."[11] And Jaegwon Kim writes: "The Cartesian model of a *bifurcated* world has been replaced by that of a *layered* world, a hierarchically stratified structure of 'levels' or 'orders' of entities and their characteristic properties."[12] As an example, he mentions the "bottom level," "consisting of whatever microphysics is going to tell us are the most basic physical particles out of which all matter is composed (electrons, neutrons, quarks, or whatever)."[13]

[10] See, e.g., Hettema (2012) and Manafu (2015).

[11] See Schaffer (2003, p. 498).

[12] This passage from Kim (1993, p. 337) is also quoted in Schaffer (2003).

[13] Ibid.

Again, there are at least two ways in which this can be made more precise:

- *The entity-based understanding*: this is based on the idea that each ontological level corresponds to a particular set of level-specific entities and perhaps their properties.
- *The fact or world-based understanding*: this is based on the idea that each ontological level corresponds to a particular set of level-specific facts and by implication a level-specific way of defining worlds.

I will ultimately endorse only the second of these understandings.[14]

3.1 Levels of entities

The entity-based way of understanding ontological levels is the most conventional one. Its key idea is that, at each level, there are certain level-specific entities, which serve as building blocks of higher-level entities. Recall, for instance, what Jaegwon Kim says about how people conventionally think about the fundamental level: it "consist[s] of whatever microphysics is going to tell us are the most basic physical particles out of which all matter is composed (electrons, neutrons, quarks, or whatever)."[15] On this understanding, higher levels consist of more complex entities, such as molecules in chemistry or cells or organisms in biology.

A version of this understanding of levels can already be found in the writings of some British Emergentists, as for instance in the following quote from C. Lloyd Morgan: "Each higher entity in the ascending series is an emergent 'complex' of many entities of lower grades, within which a new kind of relatedness gives integral unity."[16] The entity-based understanding of levels can also be found in a classic article by Paul Oppenheim and Hilary Putnam, who write: "Any thing of any level except the lowest must possess a decomposition into things belonging to the next lower level."[17] Similarly, William Wimsatt characterizes "levels of organization" as "compositional levels—hierarchical divisions of stuff (paradigmatically but not necessarily material stuff) organized by part-whole relations, in which wholes at one level function as parts at the next (and at all higher) levels, though one of the

[14] Others have drawn similar distinctions and supported the second understanding. Notably, Block (2003, pp. 141–142) contrasts "a notion of level keyed to objects" and another "keyed to relations among properties" and defends the latter, and Himmelreich (2015, Appendix B) contrasts a mereological understanding of levels and a world/state-based understanding and argues for the second. Relatedly, Norton (2014) distinguishes between different criteria for distinguishing between lower and higher levels in physics. One criterion focuses on the states of a system (distinguishing between micro- and micro-states), while the other focuses on the number of components of a system.

[15] See Kim (1993, p. 337).

[16] For this quote, see Kim (2002, p. 10).

[17] See Oppenheim and Putnam (1958, p. 9).

features of levels … is that levels are usually decomposed one level at a time, and only as needed."[18] To illustrate, he adds: "Thus, neurons are presumably composed of parts like membranes, dendrites, and synapses, which are in turn made of molecules, which are in turn made of atoms, etc., down to quarks."[19]

On such an entity-based understanding, inter-level relations are mereological relations, such as *composition* or *parthood* relations. One level is "higher" than another if the entities of the former (higher) level are composites or aggregates of the entities of the latter (lower), or conversely, the entities of the latter (lower) level are the parts or building blocks of the entities of the former (higher). Again, this would yield a partial ordering over levels.

However, as critics such as Jaegwon Kim, Angela Potochnik, and Brian McGill have pointed out, the entity-based understanding of levels has several shortcomings.[20] First, it is not clear that part-whole relationships always capture lower-versus-higher-level relationships. Only some part-whole relationships seem to do so. Plausibly, the elementary particles in physics of which larger entities are composed are associated with a lower level than, say, cells in biology. But it is not plausible, as Kim observes, that "a slab of marble is a higher entity than the smaller marble parts that make it up."[21] Similarly, Potochnik and McGill note:

> It may be that every whole is composed of smaller parts … But it is certainly not the case that every whole is composed of only parts at the next lower level. Nor is it the case that each type of whole is composed of all and only the same types of parts.[22]

Second, it is unclear that every entity can be associated with a unique level. For instance, an organism or a computer might have both physical properties and higher-level ones, such as biological or computational ones. One would then not be able to say which level the organism or computer, *qua* entity, belongs to. Is it low level, is it high level, or is it both? In particular, unless we clarify which *properties* or which *mechanisms* of the entity we are interested in, the answer seems unclear.

Potochnik and McGill point out that scientists have similar reservations about the entity-based understanding of levels, noting that "[t]he parts and wholes of the classic compositional hierarchy do not uniformly constitute nested levels of mechanisms."[23] Relatedly, Alexander Rueger and Patrick McGivern observe:

[18] See Wimsatt (1994, p. 222).
[19] Ibid., pp. 222–223.
[20] See, e.g., Kim (2002) and Potochnik and McGill (2012).
[21] See Kim (2002, p. 11).
[22] See Potochnik and McGill (2012, p. 127).
[23] Ibid., p. 132.

> When physicists talk about levels, they often do not have in mind a mereological ordering of entities. Instead, what they describe is best understood as a stratification of reality into processes or behaviours at different *scales*.[24]

It is really the mechanisms and their properties that matter from a scientific perspective rather than the entities by themselves and their mereological part-whole relations.

3.2 Levels of facts

The shortcomings of the entity-based understanding of ontological levels motivate the alternative, fact- or world-based understanding. On this understanding, it is not entities that are primarily assigned to levels but rather facts (or properties of the world). For example, some facts belong to the fundamental physical level, such as facts about the physical microstate of the universe, while other facts belong to higher levels, such as facts about metabolism in biology, mental states in psychology, or inflation and the exchange rate in economics.

If, for the moment, we run with the idea that different levels can be associated with different level-specific facts, we can see that the notion of "the world" can also be defined in a level-specific way. To introduce this idea, let's begin by recalling the standard notion of a possible world, as we find it, for instance, in Wittgenstein's *Tractatus*: "The world is everything that is the case."[25] On this picture, a *world* is a full specification of all facts that obtain at that world. Moreover, consistently with a fact-based rather than entity-based ontology, Wittgenstein emphasizes that we should think of the world as "the totality of facts, not of things."[26] Now, to incorporate the idea that facts are level-specific, we must amend Wittgenstein's definition. We can define a *possible world at a particular level* as a full specification of the way the world might be at that level. For instance, the world at the microphysical level is the totality of microphysical facts; the world at the biological level is the totality of biological facts; the world at the psychological level is the totality of psychological facts; and so on. Amending Wittgenstein's definition, we can say: "The world at a particular level is everything that is the case at that level."

The world at some higher level, under this definition, will omit certain lower-level facts—for instance, facts about certain microphysical details—that are irrelevant at the higher level. From a lower-level perspective, higher-level worlds may then look like *partial* worlds. However, from a higher-level perspective, this would be the wrong interpretation, since, as far as higher-level facts are concerned, they are complete specifications of those.

[24] See Rueger and McGivern (2010, p. 382), quoted in Potochnik and McGill (2012, p. 135).
[25] See Wittgenstein (1922, §1).
[26] Ibid., §1.1.

Higher-level worlds might also include some other facts which, despite being somehow *determined* by lower-level facts, are not explicitly included in any purely lower-level factual inventory of the world. For instance, if a certain version of non-reductive physicalism is true, psychological-level worlds may include certain mental facts which, despite being supervenient on underlying physical facts, do not themselves qualify as physical. At any rate, at each level, a *possible world at the given level* is a total specification of all level-specific facts.

On the present understanding, we can associate each level with its own level-specific set of possible worlds: the physical level is associated with the set of all possible physical-level worlds; the biological level is associated with the set of all possible biological-level worlds; and so on. Furthermore, we can think of inter-level relations as supervenience relations between facts or, more globally, between worlds at different levels. Recall that one set of facts (call it the B-facts) *supervenes* on another set of facts (call it the A-facts) if it is impossible for the former (the B-facts) to be any different without the latter (the A-facts) being different too. A standard example is the commonly assumed supervenience of chemical facts on physical facts.

Formally, on this picture, one level counts as "higher" than another if there exists a mapping from the set of worlds associated with the latter (lower) level to the set of worlds associated with the former (higher), where that mapping has the following property:

- *Surjectivity*: for each "higher-level" world, there exists at least one "lower-level" world that is mapped to it (a *lower-level realizer* of the higher-level world).

The mapping may also have a second property:

- *Many-to-one*: for at least one "higher-level world" (perhaps many), there exists more than one "lower-level" world that is mapped to it (*multiple realizability*).

These are of course standard properties of *supervenience*. Importantly, the "many-to-one" property is optional and should not be built into the definition of supervenience because we can have cases of supervenience mappings that are not many-to-one. That is: even though supervenience often goes along with multiple realizability and in many examples of supervenience relations (such as the brain-mind relation) the supervenient properties seem multiply realizable at the subvenient level, this need not always be so.

Once more, the present inter-level relation—supervenience—yields a partial but not generally total ordering over levels. Formally, in accordance with the mathematical notation for functions, we represent a supervenience relation by a function $\sigma : \Omega \rightarrow \Omega'$, where Ω (the domain of σ) is the relevant set of lower-level worlds and Ω' (the co-domain of σ) is the relevant set of higher-level worlds.

Alternatively, levels of facts, or levels of worlds, could be related to each other by grounding relations, such as when we say that the physical facts ground the chemical ones or that the chemical facts ground the biological ones.[27] But for reasons that will become clearer later, I here prefer to focus on supervenience. Importantly, neither supervenience nor grounding, which are suitable inter-level relations on a fact- or world-based understanding, should be confused with the mereological part-whole relations on the entity-based understanding.

4 A unifying framework

So far, I have reviewed four notions of levels, two of an epistemic sort and two of an ontic sort. Since we find each of these notions in some discourse about levels, does this suggest that "levels" talk is inherently diverse and pluralistic, and that there is no hope of unifying or at least reconciling all the different ways of understanding levels? Or can we find something that all these different notions have in common, and/or identify some interesting relationships between them?

What I want to show is that all four ways of thinking about levels and inter-level relations can be subsumed under a single unified framework. This framework further allows us to compare some key aspects of the different notions and to address some additional questions about levels and their relations.

I will proceed by first giving an abstract definition of a *system of levels* and then showing that each understanding of levels defines precisely such a system.[28] We can subsequently compare the systems of levels that are defined by the different understandings.

4.1 A system of levels

A *system of levels* is an ordered pair $\langle \mathcal{L}, \mathcal{M} \rangle$, defined as follows:

- $\mathcal{L}$ is a class of objects called *levels*;
- $\mathcal{M}$ is a class of directed arrows (mappings) between levels in $\mathcal{L}$, called *(inter-level) morphisms*, each of which has a *source level* L and a *target level* L' and is of the form $\mu: L \to L'$.

[27] On grounding, see, e.g., Schaffer (2009) and Rosen (2010).
[28] I first introduced this formalism in List (2019a).

A system of levels, I propose, must ideally satisfy three conditions:

(1) *Closure under composition*: if $\mathcal{M}$ contains a mapping from level L to level L' and a mapping from L' to L'', it also contains a composite mapping from L to L''.
(2) *Identity*: for each level L, $\mathcal{M}$ contains an identity mapping from L to itself.
(3) *Uniqueness*: for any pair of levels L, L', $\mathcal{M}$ contains *at most* one mapping from L to L'.

Condition (1) captures the *transitivity* of the inter-level relation encoded by the morphisms: whenever level L stands in this relation to level L', and level L' stands in this relation to level L'', then level L also stands in the relation to level L''. If the inter-level relation is supervenience, for example, then it is clearly transitive. Condition (2) captures the *reflexivity* of the inter-level relation: each level L stands in the relevant relation to itself (perhaps trivially or vacuously). In the example of supervenience, each level trivially supervenes on itself. Condition (3) captures the *uniqueness* of the inter-level relation: whenever two levels L and L' are related by it, then that relation must be unique, though two levels could well be unrelated to each other. Taking again the example of supervenience, the supervenience relation between two levels—when it exists—is clearly unique.

Mathematically speaking, the pair $\langle \mathcal{L}, \mathcal{M} \rangle$ is an algebraic structure called a "category."[29] A *category* is an ordered pair consisting of a class of objects and a class of mappings between objects ("arrows" or "morphisms") satisfying closure under composition (1) and the existence of an identity map (2). A category whose mappings additionally satisfy the uniqueness property (3) is called a "posetal category." In the present application, the "objects" are levels, and the arrows or mappings are inter-level morphisms. So, a system of levels, as formally defined here, is a special instance of a posetal category.

4.2 The four notions of levels revisited

It should be evident that all four notions of levels that I have discussed—two epistemic ones and two ontic ones—give rise to an ordered pair $\langle \mathcal{L}, \mathcal{M} \rangle$. Let's briefly run through them.

First, on the coarse-graining understanding of levels (intended to capture levels of description or levels of explanation),

- the elements of $\mathcal{L}$ are equivalence relations on some underlying set of possibilities, namely one equivalence relation for each level, and
- $\mathcal{M}$ contains precisely one inclusion mapping for every pair of equivalence relations that stand in an inclusion relation to one another.

[29] See Marquis (2015).

Second, on the linguistic understanding of levels (also intended to capture levels of description or levels of explanation),

- the elements of $\mathcal{L}$ are descriptive or explanatory languages, namely one language for each level, and
- $\mathcal{M}$ contains precisely one translation function for any pair of such languages that stand in a reducibility relation to one another.

Third, on the entity-based understanding of levels (intended to capture levels of reality),

- the elements of $\mathcal{L}$ are classes of level-specific entities, namely one class of entities for each level, and
- $\mathcal{M}$ contains precisely one arrow for any pair of such classes where the entities in one of them are parts or building blocks of the entities in the other.

Finally, on the fact- or world-based understanding of levels (also intended to capture levels of reality),

- the elements of $\mathcal{L}$ are sets of level-specific worlds, namely one set of all possible level-specific worlds for each level, and
- $\mathcal{M}$ contains precisely one supervenience mapping for any pair of levels that are related by supervenience.

It should also be evident that, with the possible exception of the entity-based understanding, all of the different understandings of levels define systems of levels satisfying the key category-theoretic conditions of (1) closure under composition, (2) identity, and (3) uniqueness. Let us first consider the inter-level relations under the first, second, and fourth understandings of levels (coarse-graining, linguistic, and fact-based), setting aside the entity-based understanding. The relevant inter-level relations, namely, the inclusion relation between equivalence relations, the reducibility relation between languages, and the supervenience relation between facts or worlds, are all transitive, as required by condition (1). Furthermore, each of these relations trivially admits identity as a special case, i.e., inclusion, reducibility, and supervenience are each reflexive, as required by condition (2). And finally, each of these inter-level relations, when it exists between two levels, is unique, as required by condition (3).

In the case of a parthood or composition relation, corresponding to the entity-based understanding, the transitivity requirement of condition (1) might still be relatively unproblematic.[30] But it is unclear that parthood or composition are

[30] For a discussion of arguments for and against the transitivity of parthood, see Varzi (2006).

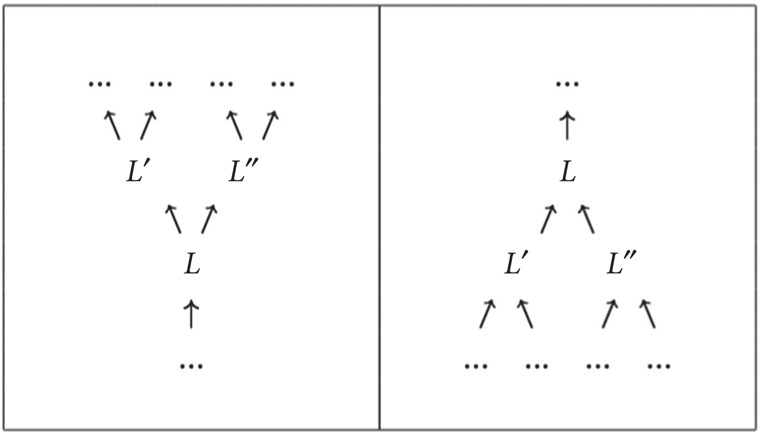

Figure 1.1 Non-linear systems of levels.

reflexive, so condition (2) may well be violated. That is, it is not obvious (though still accepted by some accounts of mereology) that a whole is a part of itself.

Similar remarks would also apply if we were to adopt a fact- or world-based understanding of levels but used grounding instead of supervenience as the inter-level relation. Grounding is not only irreflexive—no fact grounds itself—thereby violating condition (2), but its transitivity is also controversial.[31] Because of its (arguably) neater formal properties, I here prefer to use supervenience rather than grounding as the default inter-level relation for the fact- or world-based understanding of ontological levels. Still, it is worth noting that the present framework, with suitably weakened conditions on a system of levels, could also capture ontological levels that are related via grounding.

4.3 Some broader observations

All four notions of levels I have discussed share the feature that they do not generally give us a linear hierarchy of levels but just a partial ordering. This vindicates a critical remark that Jaegwon Kim made about Oppenheim and Putnam's understanding of levels: "If a comprehensive levels ontology is wanted, a tree-like structure is what we should look for; it seems to me that there is no way to build a linear system like … Oppenheim-Putnam's that will work."[32] For instance, a system of levels could look like one of the examples in Figure 1.1.[33]

[31] See Schaffer (2012).

[32] See Kim (2002, pp. 17-18).

[33] This figure is reproduced from List (2019a).

The category-theoretic perspective confirms that a linearly ordered system of levels, with a fundamental level at the bottom, is just a very special case. Levels could not only be partially rather than totally ordered, but there could also be infinitely descending chains of levels that do not terminate in any bottom level. This, in turn, shows that a "metaphysic of infinite descent," as considered by Jonathan Schaffer in his discussion of whether there is a fundamental level, is coherent, even if a "bottomless ontology" may not ultimately be supported by our best scientific theories of reality.[34] From a historical perspective, however, it is interesting to note that whenever scientists thought that they had hit "rock bottom" and identified the most fundamental building blocks of nature, new discoveries eventually led them to identify even more fine-grained constituents. Think of the move from atoms to electrons, neutrons, and protons, and subsequently to smaller elementary particles, and now to even tinier strings or superstrings of which everything may be composed.

In addition to vindicating the coherence of a non-linear and even bottomless system of levels, the category-theoretic perspective also allows us to study structural relationships between different systems of levels, including relationships between systems of levels of description on the one hand and systems of levels of reality on the other. Technically, it yields a criterion for saying when one system of levels is a *subsystem* of another, and it allows us to identify structure-preserving mappings (so-called *functors*) between different such systems, which can capture structural commonalities between them. We call one system of levels, $\langle \mathcal{L}, \mathcal{M} \rangle$, a *subsystem* of another, $\langle \mathcal{L}', \mathcal{M}' \rangle$, if $\mathcal{L}$ is a subset of $\mathcal{L}'$ ($\mathcal{L} \subseteq \mathcal{L}'$), $\mathcal{M}$ is a subset of $\mathcal{M}'$ ($\mathcal{M} \subseteq \mathcal{M}'$), and composition and identity in $\langle \mathcal{L}, \mathcal{M} \rangle$ are defined in the same way as in $\langle \mathcal{L}', \mathcal{M}' \rangle$. If one scientist thinks there are more levels than recognized by another scientist, then the system of levels according to the second scientist is a subsystem of that according to the first. A *functor* is a mapping F from one system of levels, $\langle \mathcal{L}, \mathcal{M} \rangle$, to another such system, $\langle \mathcal{L}', \mathcal{M}' \rangle$, where F assigns, to each level L in $\mathcal{L}$, a level $L' = F(L)$ in $\mathcal{L}'$, and to each mapping μ in $\mathcal{M}$, a mapping $\mu' = F(\mu)$ in $\mathcal{M}'$ such that composition of mappings and identity are preserved. In the next section, I will give one example of such a structure-preserving mapping, namely between a system of levels of description and a system of ontological levels.

5 The relationship between levels of description and levels of reality

I have raised the question of whether "levels" should be regarded mainly as an epistemic phenomenon, i.e., a feature of how we think about the world, or also as an

[34] See Schaffer (2003, p. 499). Marcus Pivato and I have also discussed such a scenario in List and Pivato (2015).

ontic phenomenon, i.e., a feature of reality itself. And I have asked how levels in the epistemic sense, which we undeniably find in science, relate to levels in the ontic sense.

In response, I will now sketch a technical and a philosophical argument for the thesis that levels of description or levels of explanation do indeed correspond to levels of reality. On the assumption that this thesis is correct, I will then consider the relationship between supervenience, which is a key inter-level relation on the ontic side, and reducibility, which is a key inter-level relation on the epistemic side.

5.1 Do levels of description or levels of explanation correspond to levels of reality?

I will first sketch a purely formal answer to this question, and I will then suggest a philosophical answer.[35] Formally, I will show that any language in the technical sense defined in Section 2.2 induces a corresponding set of level-specific possible worlds, as defined in Section 3.2. To establish this claim, let **L** be a descriptive or explanatory language. This allows us to define—at least in formal terms—a corresponding set of worlds, which we may call Ω_{L}. Specifically, we can identify the elements of Ω_{L}, the "worlds," with *maximal consistent* subsets of **L**, i.e., sets of sentences from **L** that are consistent but where the addition of *any* further sentence from **L** would introduce an inconsistency. One can think of each element of Ω_{L} as a minimally rich world that "settles" everything that can be expressed in **L**. To *settle* a sentence in **L** is to assign a truth-value to it: "true" or "false." A sentence ϕ in **L** is true at some world ω in Ω_{L} if ϕ is contained in the maximal consistent subset of **L** representing ω, and ϕ is false if it isn't. Each element of Ω_{L} thus picks out a way the world could be (a "possible world") such that everything that can be expressed in **L** is settled and nothing else is settled that isn't entailed by a set of sentences expressible in **L**. Of course, we need not literally think of a maximal consistent subset of **L** as a world, but we can think of it as representing a world. As soon as we are treating the sentences in **L** as having truth-conditions, we are thereby at least implicitly postulating the existence of some world ω in Ω_{L} that determines which sentences in **L** are true and which not.

Applying this reasoning to an entire system of levels of description or levels of explanation, we can see that each language **L** in the given system induces a corresponding set of level-specific worlds Ω_{L} within a system of ontological levels. Moreover, whenever two languages **L** and **L**′ stand in a reducibility relation, then the worlds in the corresponding sets Ω_{L} and $\Omega_{\mathrm{L}'}$ are related by supervenience. To see this, let *f* be the translation function from **L**′ (the higher-level language) to

[35] My formal answer is based on the analysis in List (2019a), but that paper did not contain an explicit formal argument to the effect that reducibility implies supervenience.

L (the lower-level language), and consider any lower-level world ω in Ω_{L}. We need to show that this determines a supervenient higher-level world ω' in $\Omega_{L'}$. Let ω' be given by the set consisting of every sentence ϕ from the higher-level language **L**′ whose lower-level counterpart $f(\phi)$ is true at ω. Since f preserves logical properties such as consistency and inconsistency, the set of higher-level sentences thus defined forms a consistent subset of **L**′. To see that it is *maximal* consistent, consider any other sentence ψ from **L**′ that is not yet included in it. It follows from the definition of our set that $f(\psi)$ is not true at ω, so its negation $\neg f(\psi)$ is true at ω. Since f preserves negation, $f(\neg\psi)$ must also be true at ω, and therefore $\neg\psi$ meets the membership criterion for the set of sentences defining ω'. We can then infer that this set together with ψ is inconsistent, and consequently that ω' is indeed maximal consistent. So, the mapping that assigns to each lower-level world ω the higher-level world ω' thus constructed qualifies as a supervenience mapping from Ω_{L} to $\Omega_{L'}$.

In this way, we have arrived at a functor which maps a given system of levels of description or levels of explanation, with reducibility as the inter-level relation, to a system of ontological levels, with supervenience as the inter-level relation.

This formal result also suggests a philosophical answer to the question of whether levels of description or levels of explanation correspond to levels of reality. It is this: we can take the fact that levels of description or levels of explanation are so useful and even indispensable in science as indicative of an underlying levelled ontology of reality. An idea along these lines already appears in Wimsatt's work. He recognizes that Ockham's razor principle is often invoked to argue for a very simple and presumably level-free ontology of the world, but he then notes:

> But Ockham's razor (or was it Ockham's eraser?) has a curiously ambiguous form—an escape clause which can turn it into a safety razor: How do we determine what is necessary? With the right standards, one could remain an Ockhamite while recognizing a world which has the rich, multi-layered, and interdependent ontology of the tropical rain forest—that is, our world.[36]

In effect, we can offer a distinctive version of the familiar "no miracles" argument to support scientific realism about ontological levels.[37] The idea is that if science supports a certain system of levels of description $\langle \mathcal{L}, \mathcal{M} \rangle$ in our best explanations

[36] See Wimsatt (1994, p. 208). For a critical perspective on this kind of argument, see, however, Heil (2003), who suggests that the tendency to infer the existence of levels of reality from our use of different levels of description or explanation stems from a problematic "Picture Theory of language according to which we can 'read off' features of the world from ways we describe the world" (p. 205). He writes: "We do not need a commitment to ontological levels to accommodate irreducible, projective predicates definitive of everyday domains and those of the special sciences. We may find it occasionally useful to speak of levels of description or explanation, but these must not be confused with levels of being or encourage the image of a layered world" (p. 220).

[37] I am grateful to the editors for suggesting the reference to the "no miracles" argument here. The "no miracles" argument famously goes back at least to Putnam (1975, especially p. 73).

of reality, then this is good evidence—a good indicator—that reality itself contains a corresponding system of ontological levels $\langle \mathcal{L}', \mathcal{M}' \rangle$. It would be surprising—a kind of "miracle"—if a level-differentiated approach to science worked so well and yet there was nothing in reality that corresponded to it.

On this picture, each level-specific language **L** in $\mathcal{L}$ picks out a corresponding ontological level in $\mathcal{L}'$, given by the set of level-specific worlds Ω_L derivable from **L**. Further, $\mathcal{M}'$ consists of supervenience mappings between appropriately related pairs of such sets of level-specific worlds. Minimally, a supervenience relation holds between any two levels for which the corresponding level-specific languages stand in a reducibility relation, but we shall see in the next subsection that reducibility is not necessary for supervenience.

The upshot is that the system of levels of description supported by science might mirror a system of ontological levels "out there in reality."

5.2 Does supervenience imply reducibility?

We have seen that whenever two distinct level-specific languages stand in a reducibility relation, then the facts or worlds at the higher one of the two corresponding ontological levels supervene on the fact or worlds at the lower. But what about the converse? Is it also true that whenever the facts or worlds at some higher level supervene on those at some lower level, then the corresponding higher-level descriptions are reducible to the relevant lower-level ones?[38]

I will now answer this question in the negative: supervenience is not sufficient for reducibility. To establish this, consider two distinct level-specific languages **L** and **L**′, and let Ω_L and $\Omega_{L'}$ be the corresponding level-specific sets of possible worlds. Moreover, suppose that there is a supervenience mapping from the lower one of the two levels to the higher, formally a surjective function σ from Ω_L to $\Omega_{L'}$. I want to show that, under plausible assumptions, the existence of a translation function for reducing the higher-level language **L**′ to its lower-level counterpart **L** is not guaranteed but rather a very special case. Recall that such a translation function, say *f*, would have to assign to each sentence in the higher-level language **L**′ an "equivalent" sentence in the lower-level language **L**, where logical properties are preserved under translation. To capture the requirement of "equivalence," in turn, we require that whenever ϕ is a higher-level sentence and *f*(ϕ) is its lower-level counterpart, the set of worlds ω in Ω_L at which the lower-level sentence *f*(ϕ) is true—call that set [*f*(ϕ)]—is the inverse image, under the supervenience mapping σ, of the set of worlds ω′ in $\Omega_{L'}$ at which the higher-level sentence ϕ is true, denoted [ϕ]. Formally,

$$[f(\phi)] = \sigma^{-1}([\phi]) = \{\omega \in \Omega_L : \sigma(\omega) \in [\phi]\}.$$

[38] The idea of supervenience without reducibility goes back at least to Fodor (1974) and Putnam (1967).

Could there be such a translation function? Suppose that

(1) the set Ω_L of lower-level worlds is infinite, in line with the assumption that infinitely many distinct initial conditions of the world are at least in principle nomologically possible;
(2) the languages we are considering, including the lower-level language **L**, are countable, in the sense that they permit the expression of as many sentences as there are natural numbers, but no more; this is a feature of practically all familiar formal and natural languages, from standard propositional logic to English.

From assumption (1), it follows that there are uncountably many subsets of Ω_L (because any infinite set has uncountably many subsets); and from assumption (2), it follows that only countably many of them are describable in the lower-level language **L**, in the sense that there exists some sentence ψ in **L** whose content $[\psi]$ matches the given subset of Ω_L (because the language admits only countably many sentences). In consequence, *almost all* subsets of Ω_L, i.e., all but a countable number, are *not* describable by a sentence (or equivalently, even by a finite logical combination of sentences) from the lower-level language **L**. This has an immediate implication for our question of whether we can assume the existence of a translation function from the higher-level language **L′** to the lower-level language **L**. Take any higher-level sentence ϕ. Given supervenience, it will certainly be true that there exists some set of lower-level worlds that forms the "supervenience base" of the content expressed by ϕ. Formally, the set $\sigma^{-1}([\phi])$ will exist and be a subset of Ω_L. However, since almost all subsets of Ω_L are not describable by any sentence from **L**, it would be a highly special case if $\sigma^{-1}([\phi])$ were so describable. Therefore, we cannot generally assume that there will exist a sentence ψ in **L** whose content $[\psi]$ is equal to $\sigma^{-1}([\phi])$. And so, the existence of a translation function *f* from **L′** to **L** is the exception rather than the rule, in combinatorial terms. I conclude that supervenience does not imply reducibility.

Of course, one could try to formulate additional conditions under which supervenience does imply reducibility. Notably, Neil Dewar, Samuel Fletcher, and Laurenz Hudetz have proposed two conditions on the two languages **L** and **L′** that are jointly sufficient for supervenience to imply reducibility.[39] One condition, called *compatibility*, requires, informally, that if the two languages share some vocabulary, they "agree" with regard to things expressible in the shared vocabulary. The other condition, called *joint characterizability*, requires, in the authors' own informal gloss, that "the union of two levels of description relative to a supervenience map admits of a description itself."[40] Now, compatibility

[39] See Dewar, Fletcher, and Hudetz (2019).
[40] Ibid.

seems to me to be a relatively undemanding condition. Moreover, it does not require the existence of any shared vocabulary between the two languages at all; it only requires that *if* there is some shared vocabulary, its meaning must be matched. Joint characterizability, however, seems much more demanding, as the authors recognize. If we take the example of psychology and fundamental physics, should we really assume that the union of these two levels of description admits a joint description itself? I take it that there is such a joint description whenever we are able to spell out explicit bridge laws between the levels in question, but often we aren't able to spell out such bridge laws. For this reason, the assumption of joint characterizability seems to me to come close to the assumption that there are explicitly describable bridge laws, in which case it is less of a surprise that this condition is favorable to the existence of a translation function between the two languages. So, I suggest that even though Dewar, Fletcher, and Hudetz have obtained an interesting formal result which may be applicable to some cases of inter-level relations, the cases it covers remain special, and we cannot generally assume that when there is supervenience, there is also reducibility.

As an aside, an analysis similar to the one given in this subsection would also show that if Ω_1 and Ω_2 are distinct ways of coarse-graining some underlying set Ω of possible worlds, representable by distinct equivalence relations on Ω, then the inclusion of the equivalence relation representing Ω_1 within the one representing Ω_2 would not imply the reducibility of a language we might use to describe Ω_1 to a language we might use to describe Ω_2. This speaks to the question of how the two epistemic inter-level relations mentioned earlier, inclusion of equivalence relations on the one hand and reducibility on the other, relate to one another.

6 Some further payoffs and applications

Arguably, many philosophical problems concern the relationship between phenomena that are intuitively at different levels, and so the present framework offers some resources for thinking about such problems. I will here mention just a few examples.

6.1 The compatibility of free will and determinism

Free-will skeptics often argue that because everything in the world is governed by the fundamental laws of physics, there is no room for free will. Humans might have the illusion that they are able to choose and control their own actions, the skeptics say, but in reality everything is determined by underlying physical processes over which we have no control.[41] One way to respond to this kind of free-will skepticism is to note that free will and choice are phenomena at the level of agency rather than at the level of physics. In particular, we can speak about free will and choice only if we use the concepts and categories of psychology and the human sciences. Without those concepts and categories, we would not be able to refer to agents and their actions, let alone ask whether these qualify as free. By contrast, the underlying physical processes, for instance those in the brain and body, are sub-agential phenomena, which belong to the level of physics, biology, or neuroscience. Many of the skeptical arguments fail to recognize the multi-levelled nature of the free-will problem and involve a mixing of levels.

To give just one example, free will plausibly requires the possibility of doing otherwise, i.e., of choosing between alternative actions, and at first sight there seems to be no such possibility if the fundamental laws of physics are deterministic, and determinism has not yet been ruled out by the physical sciences. However, once we carefully distinguish between the level of physics and the level of agency, we can see that each level is endowed with its own modal notions: possibility at the level of agency ("agential possibility") on the one hand, and physical possibility on the other. These are distinct notions, just as chemical possibility, biological possibility, and economic possibility are distinct. This insight, in turn, leaves room for showing that the possibility of doing otherwise at the level of agency can co-exist with determinism at the level of physics. Conditional on the state of the world *as specified at the level of agency*, different courses of action may be open to me and thus *agentially possible* for me, even if there is some sub-agential specification of the state of the world *at the level of microphysics* at which only a single physical trajectory is *physically possible*. There is no contradiction here: at the level of physics, we would not even be able to speak about the choices that I could or could not make; the agential "can" does not belong to the vocabulary of physics. At the level of agency, on the other hand, we would not be able to refer to, or conditionalize on, the detailed physical microstate. So, it would also make little sense to say that "conditional on the physical microstate, it is agentially impossible for me to act otherwise." This claim would mix two different levels of description that do not go together and between which there is arguably no relation of reducibility. Arguments for the incompatibility of free will and determinism, such as van Inwagen's famous consequence argument, tend to draw conclusions about what agents can and cannot do from premises about the constraints that

[41] This kind of skepticism is reviewed (with literature references) in List (2019b), which (along with List 2014) is also the source of the response summarized here. Others who have defended free will by arguing, in a variety of ways, that free will is a higher-level phenomenon rather than a physical-level one include Kenny (1978), Dennett (2003), Siderits (2008), and Carroll (2016). Furthermore, as Koons (2002) has recently pointed out and further elaborated, Wilfrid Sellars, who famously discussed the contrast between what he called the "scientific image" and the "manifest image," held a view on free will that is arguably a precursor to the one sketched here.

the fundamental laws of physics place on the physical microstate, thereby in effect conflating physical and agential levels.[42]

6.2 The level-specificity of dynamic properties

As already implicit in my brief discussion of free will, dynamic properties of a system, such as whether the system is deterministic or indeterministic, are best understood as level-relative properties. If we ask whether the world is deterministic or whether there is room for genuine randomness or some other source of indeterminism, the answer can be given only once we are clear about the level at which we are asking those questions. There might well be determinism at one level, say that of microphysics, and indeterminism at another, say that associated with some special science. The contrast between classical and statistical mechanics, where systems are conceptualized as, respectively, deterministic and probabilistic, is a case in point.[43]

Formally, if we think about each possible world as a trajectory the world might take through its state space across time (specifying in which state the world is at each point in time), then *determinism* means that any initial segment of any such trajectory up to any point in time admits only one continuation among the nomologically possible trajectories. *Indeterminism* means that some initial segment of some trajectory up to some point in time admits two or more distinct continuations among the nomologically possible trajectories: there is, at least sometimes, a “fork in the road.”

It is easy to see that if macro-level trajectories result from micro-level trajectories via some way of coarse-graining the underlying state space, such as with the help of some equivalence relation on the set of microstates, then the distinction between determinism and indeterminism is level-specific. Low-level trajectories could be deterministic while high-level trajectories could be indeterministic, or it could be the other way round.[44] As Jeremy Butterfield puts it, the micro- and macro-level dynamics of a system need not “mesh.”[45] When we move from a lower level of description to a higher one, we might see a kind of “phase transition” from deterministic to indeterministic dynamics or vice versa. Empirical considerations alone would then not allow us to settle the question of whether a particular system is deterministic or not, as Charlotte Werndl has pointed out.[46] The question

[42] See, e.g., van Inwagen (1975) and the response in List (2019c).

[43] For a discussion of coarse-graining in the move from classical to statistical mechanics, see Robertson (2020).

[44] For formal versions of this point, see Werndl (2009), Butterfield (2012), Yoshimi (2012), List (2014), and List and Pivato (2015).

[45] See Butterfield (2012). Bohmian mechanics also relies on this kind of insight. See, e.g., Goldstein (2021).

[46] On the observational indistinguishability of deterministic and indeterministic descriptions, see Werndl (2009).

receives a determinate answer only when we are clear about the level at which we are considering the system. Even a bottomless hierarchy of levels in which there is determinism at even-numbered levels and indeterminism at odd-numbered levels is coherent, albeit a somewhat contrived scenario.[47]

Similarly, one may argue that there can be "emergent" higher-level chance in a system that admits a deterministic lower-level description.[48] A necessary condition for non-trivial objective chance at a given level is merely the presence of the indeterminism at the relevant level, not the presence of indeterminism at some lower level.[49] We can thus see that, while within a given level objective chance is incompatible with determinism, across levels the incompatibility goes away: lower-level determinism is compatible with higher-level objective chance.

6.3 Indexical versus non-indexical and first-personal versus third-personal descriptions

In discussions of indexicality and subjectivity, it is often acknowledged that indexical facts cannot be derived from non-indexical ones and, similarly, that subjective facts cannot be derived from objective ones. David Lewis famously gives the following example:

> Consider the case of the two gods. They inhabit a certain possible world, and they know exactly which world it is. Therefore they know every proposition that is true at their world. Insofar as knowledge is a propositional attitude [with third-personal, non-indexical content], they are omniscient. Still I can imagine them to suffer ignorance: neither one knows which of the two he is.[50]

Each of the two gods has complete third-personal and non-indexical knowledge of the world, and yet lacks knowledge of his own position relative to the world: is he the one on the left or the one on the right, for example?

Similarly, even if we had complete information about the entire trajectory of the physical universe—from the beginning of time *ad infinitum*—we would not be able to infer from this what the present time is, i.e., the location of the "now," or at which spatial coordinates we are positioned, i.e., the location of the "here." In short, the non-indexical facts under-determine the indexical ones. This point is widely recognized in debates about the relationship between the B-theory of time, which gives us a tenseless picture of the world, and the A-theory, which gives us a

[47] See List and Pivato (2015). As the editors have pointed out to me, this scenario echoes some of David Bohm's ideas about an infinite number of levels. See, in particular, Talbot (2017).

[48] Ibid. On probability in the context of deterministic physics, see also Ismael (2009).

[49] This point is formally developed in List and Pivato (2015).

[50] See Lewis (1979, p. 520).

tensed picture.[51] The B-facts (such as whether one event happened before, at the same time as, or after another) under-determine the A-facts (such as what is happening now).

I suggest that we can think of non-indexical and indexical phenomena as residing on two different levels. Using the present framework, we can identify the non-indexical level with an ordinary set Ω of possible worlds, each of which is a total specification of all non-indexical facts, while we can identify the indexical level with a set of *centered worlds*, a set of ordered pairs consisting of a world ω in Ω and a center c within that world, which could be a spatio-temporal coordinate or a pointer to a particular individual.[52] Such centered worlds settle indexical as well as non-indexical facts, by including a center as a kind of location pointer. On this picture, the non-indexical level is the higher, more coarse-grained one, while the indexical level is the lower, more fine-grained one; different centers can be combined with the same total body of non-indexical facts. The non-supervenience of indexical facts on non-indexical ones is an immediate consequence.

Similarly, some philosophers of mind have argued that even if we were to specify the totality of third-personal facts about the world, i.e., those describable by the ordinary sciences, this would leave open the facts about first-personal experience: what it is like for conscious subjects to experience and perceive the world first-personally, or indeed whether there are any first-personal experiences at all.[53] If this is right, then the third-personal facts under-determine the first-personal ones. David Chalmers describes the challenge for a science of consciousness as follows:

> The task of a science of consciousness … is to systematically integrate two key classes of data into a scientific framework: *third-person data*, or data about behavior and brain processes, and *first-person data*, or data about subjective experience.[54]

In analogy with my brief discussion of indexicality, I suggest that we can think of first-personal and third-personal facts as residing on two different levels too.[55] We can amend the machinery of centered worlds to capture the idea that the facts of first-person experience hold only at what we may call "first-personally centered worlds," ordered pairs consisting of an ordinary third-personal world ω and a "locus of subjectivity" π, where π encodes a subject's first-person perspective on the world ω. The combination of ω and π will then determine not only all

[51] For an overview, see, e.g., Emery, Markosian, and Sullivan (2020).

[52] On centered worlds, see Quine (1969), Lewis (1979), Liao (2012), and Milano (2018).

[53] Classic discussions of this point include Nagel (1974), Jackson (1982), Levine (1983), and Chalmers (1996).

[54] See Chalmers (2004, p. 1111).

[55] I have discussed and developed this proposal in more detail in List (2023).

third-personal facts that hold at ω but also all first-personal facts that hold for the relevant subject.

Once more, we have a two-level structure. The first-personally centered level is given by the set of all possible first-personally centered worlds, and the third-personal level is given by the ordinary set Ω of all possible third-personal worlds. Just as, in the case of indexicality, the indexical level is lower (subvenient) and the non-indexical level is higher (supervenient), so the first-personally centered level is lower (subvenient) and the third-personal level is higher (supervenient).

This vindicates the claim, made by Chalmers and others, that the facts about first-personal experience do not supervene on the ordinary physical facts.[56] It further shows that there is a structural parallel between indexicality and subjectivity. Most notably, on the present picture, the much-discussed "hard problem of consciousness" is due to the fact that ordinary science only ever delivers third-personal explanations of third-personal phenomena, while the explanation of first-personal experience involves an explanandum that can only be found at a different, more richly specified level, namely the first-personally centered one.[57] The "hard problem" thus stems from the mismatch between the first-personally centered level, at which the explanandum of conscious experience is located, and the third-personal level, at which ordinary science seeks to offer an explanation.

6.4 Positive versus normative facts

A final illustrative application of the present framework concerns the relationship between positive and normative facts and the fact-value distinction. Positive facts, sometimes also just called "descriptive facts," are facts such as "H_2O consists of two hydrogen atoms and one oxygen atom," "green plants use light energy to convert water, carbon dioxide, and minerals into oxygen and certain organic compounds," and "increases in the interest rate tend to lead to decreases in inflation, other things being equal." Normative facts—if they exist, as moral realists assume—are facts such as "killing is wrong," "all humans deserve equal moral consideration," and "society ought, or ought not, to be organized in such-and-such a way." Similarly, evaluative facts—again, if they are genuine facts—are facts such as "education is good," "freedom is desirable," and "ecosystems are valuable."

Debates about moral naturalism and non-naturalism revolve around the question of how normative or evaluative facts relate to positive or descriptive ones. Do normative or evaluative facts supervene on positive or descriptive ones, or is

[56] See, in particular, Chalmers (1996). However, the present line of reasoning supports this claim in a way that is somewhat different from Chalmers's argument, by emphasizing the centeredness of the facts about first-personal experience, rather than their phenomenal character.

[57] For more on this, see List (2023).

this not the case? Moreover, if there is supervenience, is there also reducibility, in the sense that normative or evaluative discourse is translatable into positive or descriptive discourse? Or could we have a case of supervenience without reducibility? Normative or evaluative descriptions might be irreducible, even if the facts they express are, or supervene on, natural facts.

While the present framework can obviously not settle these difficult meta-ethical questions, it provides a formalism in which they can be articulated precisely. For a start, we can compare a purely positive and descriptive language with a normative or evaluative language. The latter is, in some ways, richer than the former, insofar as it includes deontic operators such as "ought" and "may" and/or evaluative predicates such "good," "bad," "desirable," and "undesirable," which are absent from the positive and descriptive language. The two languages—call them **L** and **L**′—clearly define different levels of description in the sense discussed in this chapter, and this already allows us to see precisely what it would mean to say that normative or evaluative discourse is reducible to positive or descriptive discourse: there would have to be a translation function from **L**′ to **L** which preserves content and logical properties. Moreover, the two languages, at least when taken at face value, can be thought to induce two corresponding ontological levels: one level would be given by the set of all possible worlds in a positive or descriptive sense, the other by the set of all possible worlds in some normatively or evaluatively augmented sense. A possible world in the latter set explicitly includes—in addition to ordinary positive facts—a specification of all normative or evaluative facts, while a possible world in the former set omits such facts or includes them at most implicitly, in case the hypothesis that they supervene on positive facts is true.

Elsewhere I have suggested that we could model "normatively augmented worlds" as ordered pairs consisting of an ordinary positive or descriptive world ω from some set Ω and a selection function f which assigns to each world ω a set of permissible worlds relative to ω.[58] Any ordered pair of the form $<\omega, f>$ will then be rich enough to settle not only the truth-value of all positive and descriptive sentences but also that of all sentences involving normative operators such as "ought" and "may." For instance, "it is obligatory that p" ("ought p") is true at the normatively augmented world $<\omega, f>$ if and only if p is true at *all* worlds that f deems permissible relative to ω, i.e., which are in the set $f(\omega)$. Similarly, "it is permissible that p" ("may p") is true at $<\omega,f>$ if and only if p is true at *some* worlds in $f(\omega)$.

Under this construction, there exists a many-to-one supervenience mapping from the set of all normatively augmented worlds to the set of positive or descriptive worlds. This mapping, σ, would simply map each ordered pair $<\omega,f>$ to its first component, i.e., $\sigma(<\omega,f>)=\omega$. So, the positive or descriptive level appears to be higher or more coarse-grained, while the normatively augmented level is lower

[58] See List (2019a).

or more fine-grained. This, in turn, would speak against the supervenience thesis entailed by normative naturalism and vindicate the claim that deriving an "ought" from an "is" is indeed a fallacy.[59]

However, if we could somehow show that one and only one selection function f is possible relative to each positive or descriptive world ω, then we might still be able to defend the naturalistic supervenience thesis. In this case, there would be a one-to-one correspondence between the positive or descriptive worlds and the normatively augmented ones. But at least from the perspective of logic, it is hard to see why only one selection function f should be *logically* possible for each ω. This is not the place to discuss these questions in any detail. I simply hope to have shown that the present framework allows us to look at them in a clear and systematic way.

In sum, I have reviewed several salient uses of the idea of levels, in both epistemic and ontic senses, and explained how they can all be accommodated within a unified framework. I have shown that this allows us to shed light on questions such as how levels of description or levels of explanation relate to levels of reality and whether supervenience implies reducibility. In this context, I have suggested that one might offer a kind of "no miracles" argument for a levelled ontology: the fact that levels of description or levels of explanation seem so useful and even indispensable in science may be viewed as indicative of an underlying levelled ontology of reality. Finally, I have considered some illustrative applications of this framework to a variety of philosophical problems, in the hope that they will inspire further applications as well as extensions of the framework itself.

References

Bechtel, William. 1994. Levels of Description and Explanation in Cognitive Science. *Minds and Machines* 4(1): 1–25.

Beckermann, Ansgar, Hans Flohr, and Jaegwon Kim, eds. 1992. *Emergence or Reduction? Essays on the Prospects of Nonreductive Physicalism*. Berlin: de Gruyter.

Block, Ned. 2003. Do Causal Powers Drain Away? *Philosophy and Phenomenological Research* 67(1): 133–150.

Brown, Campbell. 2014. Minding the Is-Ought Gap. *Journal of Philosophical Logic* 43(1): 53–69.

Bunge, Mario. 1960. Levels: A Semantical Preliminary. *The Review of Metaphysics* 13(3): 396–406.

Bunge, Mario. 1977. Levels and Reduction. *American Journal of Physiology* 233(3): R75–R82.

Butterfield, Jeremy. 2012. Laws, Causation and Dynamics at Different Levels. *Interface Focus* 2(1): 101–114.

Carroll, Sean M. (2016). *The Big Picture: On the Origins of Life, Meaning, and the Universe Itself*. New York: Dutton.

Chalmers, David. 1996. *The Conscious Mind*. New York: Oxford University Press.

Chalmers, David. 2004. How Can We Construct a Science of Consciousness? In Michael S. Gazzaniga, ed., *The Cognitive Neurosciences III*, 3rd ed., 1111–1120. Cambridge, MA: MIT Press.

Craver, Carl F. 2007. *Explaining the Brain*. Oxford: Oxford University Press.

[59] For a recent analysis of the is-ought gap, see also Brown (2014).

Davidson, Donald. 1973. On the Very Idea of a Conceptual Scheme. *Proceedings and Addresses of the American Philosophical Association* 47: 5–20.

Dennett, Daniel. 2003. *Freedom Evolves*. London: Penguin.

Dewar, Neil, Samuel C. Fletcher, and Laurenz Hudetz. 2019. Extending List's Levels. In Marek Kuś and Bartłomiej Skowron, eds., *Category Theory in Physics, Mathematics, and Philosophy, CTPMP 2017. Springer Proceedings in Physics*, vol. 235, 63–81. Heidelberg: Springer.

Dietrich, Franz. 2007. A Generalised Model of Judgment Aggregation. *Social Choice and Welfare* 28(4): 529–565.

Dietrich, Franz. 2018. Savage's Theorem under Changing Awareness. *Journal of Economic Theory* 176: 1–54.

Dupré, John. 1993. *The Disorder of Things: Metaphysical Foundations of the Disunity of Science*. Cambridge, MA: Harvard University Press.

Ellis, George F. R., Denis Noble, and Timothy O'Connor, eds. 2012. Top-Down Causation. Theme issue of *Interface Focus* 2(1): 1–140.

Emery, Nina, Ned Markosian, and Meghan Sullivan. 2020. Time. In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition) <https://plato.stanford.edu/archives/win2020/entries/time/>.

Eronen, Markus I., and Daniel S. Brooks. 2018. Levels of Organization in Biology. In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Spring 2018 Edition) <https://plato.stanford.edu/archives/spr2018/entries/levels-org-biology/>.

Floridi, Luciano. 2008. The Method of Levels of Abstraction. *Minds and Machines* 18(3): 303–329.

Fodor, Jerry. 1974. Special Sciences (or: The Disunity of Science as a Working Hypothesis). *Synthese* 28(2): 97–115.

Goldstein, Sheldon. 2021. Bohmian Mechanics. In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Fall 2021 Edition) <https://plato.stanford.edu/archives/fall2021/entries/qm-bohm/>.

Heil, John. 2003. Levels of Reality. *Ratio* 16(3): 205–221.

Hettema, Hinne. 2012. *Reducing Chemistry to Physics: Limits, Models, Consequences*. PhD thesis, University of Groningen.

Himmelreich, Johannes. 2015. *Agency as Difference-Making: Causal Foundations of Moral Responsibility*. Ph.D. thesis, London School of Economics <http://etheses.lse.ac.uk/3277/>.

Ismael, Jenann. 2009. Probability in Deterministic Physics. *Journal of Philosophy* 106(2): 89–108.

Jackson, Frank. 1982. Epiphenomenal Qualia. *Philosophical Quarterly* 32(127): 127–136.

Kenny, Anthony. 1978. *Freewill and Responsibility*. London: Routledge.

Kim, Jaegwon. 1993. The Nonreductivist's Troubles with Mental Causation. In *Supervenience and Mind: Selected Philosophical Essays*, 336–357. Cambridge: Cambridge University Press.

Kim, Jaegwon. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge, MA: MIT Press.

Kim, Jaegwon. 2002. The Layered Model: Metaphysical Considerations. *Philosophical Explorations* 5(1): 2–20.

Knox, Eleanor. 2016. Abstraction and its Limits: Finding Space For Novel Explanation. *Noûs* 50(1): 41–60.

Koons, Jeremy. 2022. Sellars on Compatibilism and the Consequence Argument. *Philosophical Studies* 179: 2361–2389.

Levine, Joseph. 1983. Materialism and Qualia: The Explanatory Gap. *Pacific Philosophical Quarterly* 64(4): 354–361.

Lewis, David. 1979. Attitudes De Dicto and De Se. *The Philosophical Review* 88(4): 513–543.

Lewis, David. 1988. Relevant Implication. *Theoria* 54(3): 161–174.

Liao, Shen-yi. 2012. What Are Centered Worlds? *The Philosophical Quarterly* 62(247): 294–316.

List, Christian. 2014. Free Will, Determinism, and the Possibility of Doing Otherwise. *Noûs* 48(1): 156–178.

List, Christian. 2019a. Levels: Descriptive, Explanatory, and Ontological. *Noûs* 53(4): 852–883.

List, Christian. 2019b. *Why Free Will is Real*. Cambridge, MA: Harvard University Press.

List, Christian. 2019c. What's Wrong with the Consequence Argument: A Compatibilist Libertarian Response. *Proceedings of the Aristotelian Society* 119(3): 253–274.

List, Christian. 2023. The Many-Worlds Theory of Consciousness. *Noûs* 57(2): 316–340.

List, Christian, and Marcus Pivato. 2015. Emergent Chance. *The Philosophical Review* 124(1): 119–152.

Manafu, Alexandru. 2015. A Novel Approach to Emergence in Chemistry. In Eric Scerri and Lee McIntyre, eds., *Philosophy of Chemistry: Growth of a New Discipline*, 39–55. Heidelberg: Springer.

Marquis, Jean-Pierre. 2015. Category Theory. In Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* (Winter 2015 Edition) <http://plato.stanford.edu/archives/win2015/entries/category-theory/>.

Milano, Silvia. 2018. *De Se Beliefs and Centred Uncertainty*. Ph.D. thesis, London School of Economics <https://doi.org/10.21953/lse.b2wsi6xbghk6>.

Modica, Salvatore, and Aldo Rustichini. 1999. Unawareness and Partitional Information Structures. *Games and Economic Behavior* 27(2): 265–298.

Nagel, Thomas. 1974. What Is It Like To Be a Bat? *The Philosophical Review* 83(4): 435–450.

Norton, John. 2014. Infinite Idealizations. In *European Philosophy of Science: Philosophy of Science in Europe and the Viennese Heritage*, Vienna Circle Institute Yearbook, vol. 17, 197–210. Dordrecht/Heidelberg/London/New York: Springer.

Oppenheim, Paul, and Hilary Putnam. 1958. Unity of Science as a Working Hypothesis. *Minnesota Studies in the Philosophy of Science* 2: 3–36.

Owens, David. 1989. Levels of Explanation. *Mind* 98(389): 59–79.

Potochnik, Angela, and Brian McGill. 2012. The Limitations of Hierarchical Organization. *Philosophy of Science* 79(1): 120–140.

Putnam, Hilary. 1967. Psychological Predicates. In W. H. Capitan and D. D. Merrill, eds., *Art, Mind, and Religion*, 37–48. Pittsburgh: University of Pittsburgh Press.

Putnam, Hilary. 1975. *Mathematics, Matter and Method*. Cambridge: Cambridge University Press.

Quine, W. V. 1969. Propositional Objects. In *Ontological Relativity and Other Essays*, 139–160. New York: Columbia University Press.

Robertson, Katie. 2020. Asymmetry, Abstraction, and Autonomy: Justifying Coarse-Graining in Statistical Mechanics. *British Journal for the Philosophy of Science* 71(2): 547–579.

Rosen, Gideon. 2010. Metaphysical Dependence: Grounding and Reduction. In Bob Hale and Aviv Hoffmann, eds., *Modality: Metaphysics, Logic, and Epistemology*, 109–135. Oxford: Oxford University Press.

Rueger, Alexander, and Patrick McGivern. 2010. Hierarchies and Levels of Reality. *Synthese* 176(3): 379–397.

Schaffer, Jonathan. 2003. Is There a Fundamental Level? *Noûs* 37(3): 498–517.

Schaffer, Jonathan. 2009. On What Grounds What. In David Chalmers, David Manley, and Ryan Wasserman, eds., *Metametaphysics: New Essays on the Foundations of Ontology*, 347–383. Oxford: Oxford University Press.

Schaffer, Jonathan. 2012. Grounding, Transitivity, and Contrastivity. In Fabrice Correia and Benjamin Schnieder, eds., *Metaphysical Grounding: Understanding the Structure of Reality*, 122–138. Cambridge: Cambridge University Press.

Siderits, Mark. 2008. Paleo-Compatibilism and Buddhist Reductionism. *Sophia* 47(1): 29–42.

Talbot, Chris, ed. 2017. *David Bohm: Causality and Chance, Letters to Three Women*. Cham: Springer.

Van Inwagen, Peter. 1975. The Incompatibility of Free Will and Determinism. *Philosophical Studies* 27(3): 185–199.

Varzi, Achille C. 2006. A Note on the Transitivity of Parthood. *Applied Ontology* 1(2): 141–146.

Werndl, Charlotte. 2009. Are Deterministic Descriptions and Indeterministic Descriptions Observationally Equivalent? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 40(3): 232–242.

Wilson, Mark. 2010. Mixed-Level Explanation. *Philosophy of Science* 77(5): 933–946.

Wimsatt, William C. 1994. The Ontology of Complex Systems: Levels of Organization, Perspectives, and Causal Thickets. *Canadian Journal of Philosophy Supplementary Volume* 20: 207–274.

Wittgenstein, Ludwig. 1922. *Tractatus Logico-Philosophicus*. London: Kegan Paul.

Yoshimi, Jeffrey. 2012. Supervenience, Dynamical Systems Theory, and Non-Reductive Physicalism. *British Journal for the Philosophy of Science* 63(2): 373–398.

2
Antireductionism Has Outgrown Levels

Angela Potochnik

"Entities at different levels"; "explanations in higher-level terms"; "the fundamental level"; "higher-level sciences." These and many similar turns of phrase are used throughout philosophy of science and metaphysics, typically without much in the way of explication. These are used as starting points for discussions rather than the endpoints of argument. By what right are such turns of phrase used? Best I can tell, the rationale for such levels talk is taken to stem from (variously) the mere fact of productive scientific inquiry addressing objects larger than fundamental particles; that scientists of various stripes invoke levels on a regular basis; that some scientific investigations target the very smallest happenings in our world, happenings that seem bound up in one way or another with everything that goes on in our world; that philosophers, scientists, and laypeople alike often have whole discourses without making reference to these smallest goings-on; that there are sometimes multiple candidate explanations of a single explanandum, some of which feature larger entities and their properties than others.

I have done the same. In a 2010 paper on levels of explanation, I simply say, "In general, a lower-level explanation cites properties of objects that stand in a part-whole relationship to objects referenced in the competing higher-level explanation" (Potochnik, 2010, p. 64), and I reference lots of important philosophers who talked about scientific explanations in this way. Since then, I have started to examine this assumption that explanations come in levels more carefully and to attend to others who are also questioning this. I have been startled at how little weight these turns of phrase and the assumption behind them can actually bear. And yet, the assumption that explanations come in levels persists as an unexamined starting point of philosophical treatments of explanation. Levels of explanation receive plenty of discussion, but the discussion largely consists in whether there are higher-level explanations and, if so, what relationship they bear to what we know, or might someday know, about the smallest, microphysical happenings in our world. As far as I can tell, that scientific explanations and the entities featured in them are arranged in levels mostly still goes uncontested.

In this paper, I will argue that it is a mistake to invoke levels in discussions of scientific explanation. The invocation of levels played a very important role historically in philosophy of science, as a way to motivate an antireductionist stance about scientific explanation. But our scientific and philosophical understanding

Angela Potochnik, *Antireductionism Has Outgrown Levels* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Angela Potochnik 2024. DOI: 10.1093/oso/9780192862945.003.0003

has progressed mightily since then, and we can do antireductionism better. It is thus time for philosophers of science to abandon the levels framework in our discussions of scientific explanation.

In Section 1, I outline the role invocation of levels has played in philosophy of science, focusing especially on how they have been used to motivate antireductionism about scientific explanations. In Section 2, I argue that framing antireductionism about scientific explanation as a thesis about levels of explanation has led to problematic commitments—that candidate explanations form a linear or at least partial hierarchy, can be ordered by generality, and bear straightforward metaphysical relationships to one another. In Section 3, I use the difficulties of the levels framing to show how antireductionism can be done better without levels. This involves reconsidering the relationships different explanations bear to one another, recognizing a wider variety of candidate explanations, and appreciating how considerations guiding the selection of explanations can vary across research projects. Finally, in Section 4, I conclude by offering a new "working hypothesis" about the nature of our scientific explanations: they are many and varied, often featuring large-scale, distant, and structural factors. The decision of explanatory quality is not about how fine-grained our characterization of local factors should be but rather which factors at what scales we should attend to. Reductionism has failed, but so too has the framework of explanatory levels. The levels framing is no longer necessary nor helpful in motivating antireductionism about scientific explanation.

1 Levels in Antireductionism

There is tradition in philosophy, as well as in at least some fields of science, to invoke levels on both sides of debates about reductionism. In philosophy, this tradition traces back at least to Oppenheim and Putnam's influential motivation for the unity of science understood as reduction to physics.[1] Oppenheim and Putnam's levels are, in descending order: social groups, multicellular living things, cells, molecules, atoms, and elementary particles. The relation among entities at different levels is one of part-whole composition.

[1] Hempel and Oppenheim (1948) also consider levels of explanation, but their levels of explanation are not compositionally defined but defined in terms of abstractness: "higher levels [of explanation] require the use of more or less abstract theoretical constructs which function in the context of some comprehensive theory" (p. 147). Their illustration is explaining a planet's position with reference to Kepler's laws (lower level) or instead from the general law of gravitation and laws of motion (higher level). Indeed, a strategy of high-level explanation on which Hempel and Oppenheim focus is "explaining a class of phenomena by means of a theory concerning their micro-structure," so the usage of "levels" is very different from the later Oppenheim and Putnam paper. Thus, although the connection between high-level explanation and greater abstractness that has been influential is established in this paper, the relationship of levels to explanatory reductionism is reversed from what is customary in later philosophical discussions.

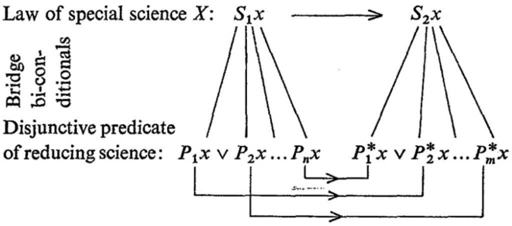
Law of special science X: S_1x → S_2x
Bridge bi-conditionals
Disjunctive predicate of reducing science: $P_1x \vee P_2x \ldots P_nx$ $P_1^*x \vee P_2^*x \ldots P_m^*x$

Figure 2.1 Fodor's illustration of why reduction of scientific explanations is unlikely to come to pass. This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

> They say: Any whole which possesses a decomposition into parts all of which are on a given level, will be counted as also belonging to that level. Thus each level includes all higher levels. However, the highest level to which a thing belongs will be considered the "proper" level of that thing. (1958, pp. 9, 10)

Note that this is not an endpoint of their analysis but rather the starting point. That is, Oppenheim and Putnam presume that this is how our world is ordered—as wholes entirely decomposable into parts occupied at a lower level—and then investigate what relation we should expect among the fields of science that investigate these levels. They predict, based on empirical evidence of how science seemed to them to be proceeding, that all science would eventually be reduced to microphysics—loosely put, that our best scientific laws would be vindicated and analyzable in terms of microphysical laws.[2]

A decade and a half later, Fodor (1974) responds directly to Oppenheim and Putnam's "working hypothesis" of the unity of science with the opposed hypothesis of disunity, i.e., the failure to reduce explanations or theories to physical theory. His argument rests on the observation that there is not often a neat relationship between kinds invoked in higher-level explanations and the physical kinds upon which they depend: "interesting generalizations can often be made about events whose physical descriptions have nothing in common" (p. 103). Thus enters the influential idea of multiple realization, and with it the presumption that high-level properties are realized by lower-level properties. We can also credit Fodor with the diagram shown in Figure 2.1, variants of which have proliferated ever since in discussions of the significance of realization and multiple realization for explanation and causation. Around the same time, Putnam (1975) emphasized the value of the generality of higher-level explanations compared to lower-level explanations, deploying the now classic example of explaining why a square peg fails to

[2] For discussion of a very different and largely neglected tradition of the unity of science tracing back to the Vienna Circle, see Potochnik (2011).

go through a round hole with the same diameter. And Garfinkel (1981) argued that explanation "seeks its own level," that the factors that truly make a difference to some occurrence are found at the same level. For Garfinkel, reductive explanations thus suggested sensitivity to details that were in fact irrelevant.

The idea that higher-level explanations are more general or more abstract than lower-level explanations has since become very influential. Often the justification given is in terms of multiple realization, as in Fodor's influential argument and diagram. Yablo (1992) employs this idea in his proportionality argument for mental causation. Sober (1999) employs the framework of multiple realization giving rise to different levels of explanation to support a pluralism about explanatory strategies, including lower- and higher-level explanations that are, respectively, more specific and more general. Jackson and Pettit (1992) deploy a different approach from Sober's to defend a pluralism that admits both more general higher-level explanations and more specific lower-level explanations. Hauge (2011) and Clarke (2016) each analyze what specific variety of abstractness might be at play in distinguishing high-level from lower-level explanations.

This combination of ideas has become a general setup for antireductionism about scientific explanation: candidate explanations come in levels; entities that feature in lower-level explanations compose the entities that feature in higher-level explanations; properties cited in lower-level explanations determine and multiply realize the properties cited in higher-level explanations; higher-level explanations are more general than lower-level explanations. Arguments in favor of mental causation (e.g., Yablo, 1992) and of metaphysical emergence (e.g., Wilson, 2013) defend not just the explanatory but the causal autonomy of higher-level properties conceived in this way. Mechanistic accounts of explanation have a conception of mechanisms consistent with this general setup and deploy it as a competing view to explanatory reductionism (e.g., Craver, 2007).

Discussions of complexity as a bulwark against reductionism also presume this general setup. Here is Herbert Simon, in his classic discussion of complexity and systems theory:

> The central theme that runs through my remarks is that complexity frequently takes the form of hierarchy, and that hierarchic systems have some common properties that are independent of their specific content. Hierarchy, I shall argue, is one of the central structural schemes that the architect of complexity uses.
>
> By a hierarchic system, or hierarchy, I mean a system that is composed of interrelated subsystems, each of the latter being, in turn, hierarchic in structure until we reach some lowest level of elementary subsystem. (1962, 468)

William Wimsatt propounded this style of view of levels in philosophy of science, famously describing levels as "local maxima of regularity and predictability"

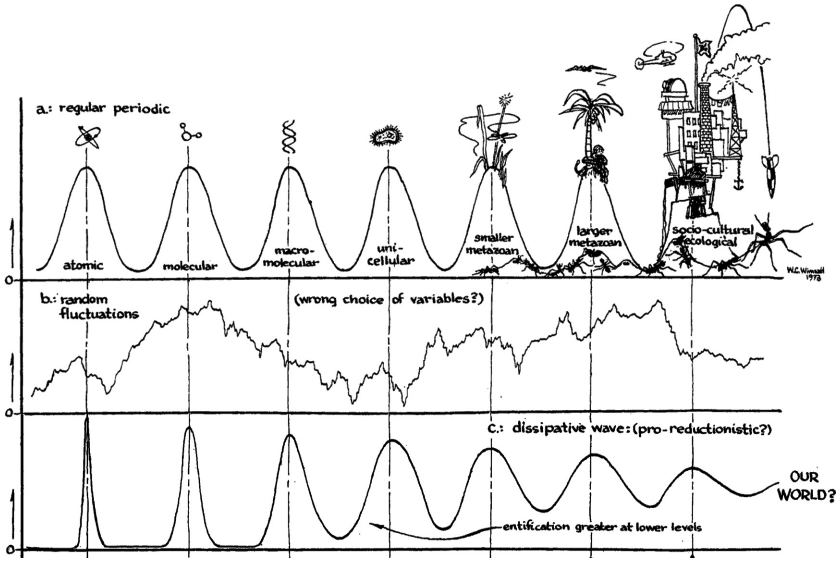

Figure 2.2 Wimsatt's illustration of levels of organization. This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

(1972; 2007). This version of antireductionism has differences from the philosophical tradition I surveyed above, but it also resonated with that tradition and has regularly been treated as an allied position. Wimsatt has his own diagram, one that has been more influential in some circles than Fodor's illustration of multiple realization; see Figure 2.2.

And so the idea that explanations come in levels, as perhaps also do causes and organizational relationships, has become entrenched as a key assumption of antireductionism about scientific explanation. My survey here has traversed a few different debates and traditions, and it is certainly incomplete. Across the philosophical discussions of levels, several related concepts of levels are variously at play, and often several are invoked without carefully distinguishing among them.[3] There is one important distinction in conceptions of levels I want to point out at this juncture: some are metaphysical, while others are representational. That is, some conceptions of levels regard the entities, properties, or processes in our world—as with claims about levels of organization or high-level causation—whereas others regard our representations of our world—as with claims about the relative abstractness of levels or the levels of our scientific theories. Claims about

[3] See Potochnik (2017, Ch. 6; 2021) for more on the variety of levels concepts and how these have been conflated.

level relationships among fields of science or of scientific laws seem to lurk somewhere in between metaphysical and representational commitments.[4]

Here is an attempt at a generic formulation that perhaps can accommodate all of this variety in the invocation of levels. Most broadly, the antireductionist position about levels of explanation seems to presume that (1) the world (or our representations thereof) is organized into levels, such that (2) our candidate explanations are structured in terms of levels, which motivates (3) the persistence of different fields of science addressing these various levels. This seems to give full voice to the antireductionist impulse, even if my characterization is abstract and rough in order to lump rather than split.

This style of argument for antireductionism has also shown up in philosophical debates about specific scientific investigations, including at least physics, biology, psychology, cognitive science, and the social sciences. The ultimate concerns of these debates vary. In psychology and cognitive science, at issue is whether there is room for mental states as explanations or even as real, causally efficacious components of our world. In physics and biology, the question is how different approaches relate to one another and to other fields entirely, chiefly fundamental physics. In the social sciences, the chief question seems to be social explanations, i.e., explanations citing entities larger than individual agents. Across all these fields, the invocation of levels to counter reductionism also seems to be in service to a stance on methodology or proper modeling approach. Even when what is at stake has varied, these debates have had remarkably similar contours.

2 What the Levels Framework Gets Wrong about Explanation

The previous section summarized how a variety of forms of antireductionism about explanation, developed in a variety of philosophical contexts, resist explanatory reductionism by invoking levels. This might seem like an obvious strategy. One might think that all that's needed from the invocation of levels in a project of antireductionism is a way to gesture at the idea that there is something other than the most fundamental: that there is more going on in the world than just microphysical happenings; that our scientific enterprise involves more than fundamental physics; that our scientific laws encompass more than fundamental physical laws; that our scientific explanations come in more varieties than microphysical explanations. But, when one puts this point in terms of levels, much more

[4] A second difference in levels conceptions that can cause confusion is whether (metaphysical) levels are a relation among types or between types and tokens. In this discussion, I presume the former. Even if there is a type-token basis for levels, each token is of many types—i.e., can be categorized with the use of multiple different descriptions—and levels have often been invoked to describe the relationships among those types. Additionally, all fields of science, including fundamental physics, target types rather than tokens.

is taken on board than just this. And, I will suggest in this section, what is taken on board is philosophically problematic. Framing debates about explanation in terms of levels of explanation systematically misconstrues the relationship different explanations bear to one another.

To get at this point, let's start by taking one step back from the levels framework. What is not controversial is that potential explanations come in *varieties*. A variety of physical and chemical theories and models bear on the behavior of gases; genes are investigated with different methods in several subfields of biology; and behavioral phenomena are targeted in studies ranging from neuroscience and molecular genetics to ecological psychology and sociology. Different scientific projects that target the same phenomena generate varieties of potential explanations. I say "potential explanations" in order to remain neutral on the question of whether all succeed as explanations. Indeed, reductionism is the view that some of these varieties of potential explanations are universally privileged over the others—or will be so at a future stage of science. Antireductionism is the rejection of that universal privilege.

Characterizing varieties of potential explanations as *levels* of explanation entails that the varieties occur in a linear hierarchy or, at the very least, a partial hierarchy. Accounts of levels of explanation tend to presume that potential explanations for some explanandum are either lower level than, higher level than, or at the same level as any other potential explanations (for that explanandum). This just is the idea that explanations are arranged in levels. It is possible, though, that levels of explanation comprise a partial hierarchy: that some potential explanations for a given explanandum are at incomparable levels, while others are related by lower-level-than, higher-level-than, or same-level-as. Invocations of levels of explanation also tend to presume, in line with Fodor's diagram (Figure 2.1), that there is a many-one relationship among explanations at different levels. Higher-level explanations have often been taken to be more general, i.e., to apply to explananda across a greater range of circumstances, and lower-level explanations to be more specific, i.e., to apply across a more limited range of circumstances (see, e.g., Putnam, 1975; Garfinkel, 1981; Jackson and Pettit, 1992; Sober, 1999). The entities or properties referenced in different levels of explanation are also supposed to bear special relationships to one another: parts and wholes (e.g., Putnam, 1975), subcomponents and components of mechanisms (e.g., Craver and Bechtel, 2007), determination (e.g., Yablo, 1992), supervenience (e.g., Kim, 1998), or realization (e.g., Fodor, 1974). Across this variety, it seems higher-level explanations are supposed to reference something that (in one way or another) depends upon what's referenced in lower-level explanations.

But the varieties of potential explanations for a given explanandum are seldom if ever arranged in any of these ways. Consider the three commitments described just above in reverse order. Do varieties of potential explanations reference properties or entities that are connected by metaphysical dependence relations (e.g., composition, determination, supervenience, or realization)? Examples philosophers appeal to in debates about explanatory reductionism tend to have this form. Fodor (1974) contrasts Gresham's Law governing monetary exchanges with imagined lower-level explanations of exchanges of (separately) wampum, dollar bills, and a signed check. Putnam (1975) contrasts the geometric explanation for the square peg not going through a round hole with an explanation in terms of the individual atomic structure of the peg and of the edges of the hole. Garfinkel (1981) contrasts the Lotka-Volterra account of the increasing fox population with an explanation in terms of individual fox and hare births and predation events. An exchange of something valued as currency is required for any monetary exchange; precise atomic structure dictates size relations; individual births and deaths combine to determine population growth. But something is a bit fishy about the lower-level candidate explanations in all of these examples. Namely, these potential explanations are not actually types of explanation generated in scientific research. Where is the scientific research on wampum exchanges (or signed-check exchanges), on the precise atomic structure of a specific one-inch-sided peg, or on the births and deaths of foxes and hares in a specific population? Without the existence of such research, we can't expect these examples to inform our judgments on how to choose among a variety of potential explanations.

In contrast, in situ examples of potential explanations for a given explanandum do not tend to have this form. Behavioral phenomena—say, a tendency to heightened aggression—are investigated in a range of different fields (Longino, 2013). Neuroscience identifies neural structures and pathways associated with this tendency; molecular genetics identifies genes associated with the tendency; ecological psychology and sociology identify social and environmental influences. None of these explanations for heightened aggression is related to another by composition, determination, supervenience, or realization. Rather, these different explanatory strategies specify different causal influences on the behavioral phenomenon in question (which may or may not also bear causal relationships to one another). Genes can causally influence neural structures and pathways but do not compose or realize them, while social and environmental influences are separate influences that can also have neural effects. This result is easy to replicate with other instances of multiple investigations targeting the same phenomenon. Potochnik (2010) argues this is the general form of the relationship between competing "levels" of explanation in science, as different potential explanations broadly fail to be related by metaphysical determination.[5] Ylikoski (2014) argues that, in the social sciences, there is a variety of micro- and macro- social explanations that are not arranged in levels.

[5] Franklin-Hall (2016) calls this relationship "horizontal" rather than "vertical"; terminology that reinforces an expectation of dependence of higher-level wholes on their lower-level parts.

At first glance, it seems genetics might fare better, with the anticipated dependence relations connecting explanations in molecular genetics to those in classical genetics. These investigations invoke different specifications (structural and functional, roughly) of the very same entities (genes). But the styles of explanations formulated in molecular genetics and classical genetics do not capitalize on that relationship. Rather, molecular genetics provides information about molecular genes associated with some trait of interest, often via genome-wide association studies. Very seldom is the causal role of any given molecular gene able to be identified. And classical genetic explanations, as in behavioral genetics, partition overall influence on a trait into genetic heritability vs. environmental influences (Longino, 2013). So, despite the apparent promise, molecular genetics and classical genetics are not well positioned to provide candidate explanations citing entities or properties related by metaphysical dependence. The upshot of this discussion is that entities and properties featured in a variety of potential scientific explanations for the same or related phenomena in scientific research generally bear no special metaphysical dependence relationship (such as composition, realization, determination, or supervenience) to one another.

This finding also interferes with the expectation that higher-level explanations are more general than lower-level explanations, as that expectation of relative generality issues from the anticipation of a many-one relationship between the relevant entities or properties stemming from multiple realization, supervenience, determination, or composition. There is also a deeper problem beyond this. The generality of an explanation depends on its degree of abstractness, i.e., how many details are specified. Specifying fewer details results in a more general account, while incorporating more details decreases the generality (and increases the precision) of an account. But degree of abstractness is a property of representations, or characterizations, not of what is being represented or characterized. Generality, as in scope of applicability, might be a metaphysical property, but the relative generality of an explanation is influenced by representational choices, namely what to include or exclude from the explanation.

Two implications of this are important for present purposes. First, the abstractness or generality of an explanation needn't relate to the metaphysical dependence relations typically thought to characterize levels, such as realization, supervenience, or composition (Potochnik, 2021).[6] By specifying additional properties or omitting mention of other factors (e.g.), one can make an explanation more or less general regardless of what entities and properties it features. Some explanations in microphysics are highly general, while others apply only in a finely

[6] The determination relation seems more closely related to generality but also more distantly related to levels of explanation as they have often been understood. Franklin-Hall (2016) acknowledges this by pointing out the divergence between determinate/determinable and micro/macro yet persists in calling the determinate/determinable relation "vertical"—which to my mind continues the conflation of fineness of specification and compositional determination that I aim to disambiguate here.

specified set of conditions. The same goes for biology and economics. Discussions of levels of explanation have generally presumed that lower level is more specific and higher level more general, but in other contexts philosophers have regularly touted the relative generality of accounts in physics vs. the so-called special sciences. Thus, the explanatory value of generality is not necessarily a point in favor of antireduction about explanation, and specifying additional explanatory detail need not result in a reductive explanation.

The second implication of abstractness being representational that I want to emphasize is that it is quite common to have *incommensurate* degrees of generality—i.e., two representations that cannot be ranked in their generality but simply are general in different respects. Abstractness (and thus generality) is achieved by omitting details. Omit or include details about different things, and the resulting representations are of incommensurate generality. They specify different aspects of the world in virtue of what is depicted and generalize to different ranges of circumstances in virtue of what is omitted. Philosophers have debated the proper degree and variety of generality in our scientific explanations, but that is incidental to the present point. The point here is simply that varieties of potential explanations quite often cannot be ranked by degree of generality, so this feature of the levels of explanation framing fails to obtain on a regular basis.

Finally, let's consider the very basis of framing varieties of potential explanations as *levels* of explanation: whether potential explanations for a given explanandum are arranged in a linear hierarchy or partial hierarchy. The delineation of levels requires that potential explanations be sortable into lower-level-than, higher-level-than, or same-level-as; if one anticipates a partial ordering rather than linear hierarchy, a fourth category of "incommensurate level" is also available. The argument above that potential scientific explanations of the same phenomenon often are not related by citing entities or properties related straightforwardly by composition, realization, supervenience, or determination already suggests difficulties with sorting potential explanations into a linear hierarchy, as those are typically cited as the basis for delineating levels. Even if one aims for a partial ordering, the enormity of the "incommensurate level" category is troubling, if—as I have argued—most or all potential explanations for the same explanandum are not related by composition, realization, supervenience, or determination. Appeal to levels of explanation was meant to categorize potential explanations in an informative way, but, for many of our candidate scientific explanations, even with the same explanandum, the anticipated means for sorting into levels are unavailable.

The problem goes deeper. I have skirted this issue so far in this chapter, but delineation of levels of explanation has typically relied on delineating levels of organization based on relations like material and mechanistic composition, spatial and temporal scale, and realization. But, as I have explored elsewhere, it turns out that these relations together do not determine a linear hierarchy or useful partial ordering (Potochnik, 2017, Ch. 6), and any individual relation among them cannot be used separately to determine a linear hierarchy or useful partial ordering (Potochnik, 2021). I want to emphasize that the problem is not that there are occasional exceptions to an ordering or orderings that cannot be universal in scope. Rather, it seems our world—or at least the properties and entities into which science has carved it—simply is not composed of levels (see also Thomasson, 2014; Eronen, 2015). Indeed, when Kim (2002)—a philosopher perhaps best known for a causal exclusion argument that presumes the levels framework—explored the basis for levels, even he came up short.

The commitments I take the levels framework to have and the difficulties I have pointed out with each of these commitments are summarized in Table 2.1. If levels of organization cannot be used to impose a linear hierarchy on our explanations, and generality rankings of our varieties of explanation do not result in a linear hierarchy, and the varieties of potential scientific explanations we observe in situ do not seem to bear any of the anticipated hierarchical relationships to one another, then I am not sure what the basis would be for the presumption that scientific explanations come in levels.

It follows from all of this that we should pause to consider the proper framing before employing the seemingly straightforward diagrams of level relations that recur in many discussions of levels of explanation and causation. Figure 2.1 depicts an important example of such a diagram; another primary example is the diagram commonly used to depict the causal exclusion argument that originated with Kim (e.g., 1998). The vertical lines illustrating realization or supervenience may seem to be uncontroversial in light of the broad acceptance of physicalism and material composition, but these commitments do not suffice as grounds for the assertion that our scientific explanations or the causal relationships they feature bear the implied metaphysical relationships to one another. Those vertical lines must be earned rather than assumed, lest the very framing of the question inherit our mistaken assumptions about the variety of potential scientific explanations.

Table 2.1 Commitments of the view that potential explanations come in levels and the difficulties facing them.

Commitments of the view that explanations come in levels	Difficulties with these commitments
Explanation varieties arranged in linear hierarchy	Incommensurate rankings are commonplace
Higher-level explanations are more general than lower-level explanations	Abstractness is a representational, not metaphysical, property
Higher-level explanations cite something that depends on what lower-level explanations cite	Potential explanations often bear no special relationship to one another

Philosophical discussions about explanatory reductionism have, by and large, presumed explanations come in levels, identified candidate explanations on the basis of that expectation, and then assessed quality and status of those candidate explanations to make a determination regarding reductionism. The debate is transformed if instead we look to science to see what varieties of potential explanations for the same phenomena are identified and what relationship those bear to one another. One favoring the standard strategy might argue that that approach gets at a metaphysically deeper picture of alternative explanations or that, in the fullness of time, varieties of scientific explanation will tend toward the predicted relationship of levels. In response to the former, I'll point out that so long as the target of our philosophical accounts of explanation is *scientific* explanations, the better strategy is one that applies to explanations actually formulated in science. In response to the latter, I see no reason to expect that science is moving toward a division of labor ordered by metaphysical relationships like composition, realization, or supervenience. This is supported by considerations like those I have already raised in this section with framing antireductionism about explanation in terms of levels.

3 Antireductionism Without Levels

Conceptualizing antireductionism in terms of levels systematically misconstrues the relationship different candidate explanations bear to one another. Candidate explanations do seem to relate many-one to their explananda, as anticipated with the levels framing. Potential explanations come in varieties. But, as discussed in the previous section, "levels" is not an apt description for those varieties. The relationship potential explanations of the same phenomena bear to one another is not a linear or partial hierarchy orderable by generality and defined by metaphysical dependence of the featured entities and properties. In this section, I use the shortcomings of the levels framing to inspire an alternative approach to antireductionism. This alternative approach better describes the variety of candidate explanations we see in scientific research and the relationships these explanations bear to one another, and it also better accounts for the considerations that guide the selection among candidate explanations. I discuss these three significances below. I conclude this section by pointing to some problematic downstream implications to which framing antireductionism in terms of levels has given rise and, accordingly, an antireductionism without levels helps us avoid.

First, having jettisoned the expectations that accompanied the levels framing, let's reconsider how different potential explanations of the same explananda relate to one another. In Section 2, I anticipated an alternative to straightforward metaphysical dependence (whether composition, determination, supervenience, or realization): that different explanatory strategies specify different causal influences on the phenomenon in question, influences that may or may not also bear causal relationships to one another (see also Potochnik, 2010).[7] Just as attention to in situ varieties of scientific explanations reveals that the levels framing is frequently inapt, this also lends prima facie support to this alternative framing. Consider again the variety of investigations that aim to explain human behavioral tendencies. These can feature (at least) molecular genes, neurological features, environmental influences, social context, and more. These factors interact in their influence on human behavior, and some also causally influence one another, as with molecular genes' and the environment's impact on neurological development.

This example supports an additional consideration in favor of the expectation of different potential explanations targeting distinct influences on a phenomenon. In a wide variety of scientific research, factors that are non-local turn out to be key influences on phenomena. This amounts to an empirical vindication of at least one form of antireductionism: it turns out that large-scale influences, distal influences, and structural influences regularly shape the happenings in our world. Examples are easy to generate. In ecology, abundance (i.e., population sizes) traditionally was thought to be determined locally by interactions with competitors but is now recognized to be shaped globally, such as in the evolution of specialists and generalists (e.g., Gaston and Blackburn, 2000). Dynamical systems theory has been fruitfully applied to research ranging from physics to ecology, cancer, and cognitive science. And it is now widely appreciated that mitigating racism involves not just changing minds but renovating social systems. The potential for significant non-local influence means there are more places to look for explanations and less reason to think independently generated explanations bear any metaphysically deep relationships to one another, such as determinables and determinates or realized and realizers. Phenomena in our world are shaped by so many different influences, operating at different timescales and spatial scales, that the potential explanations for one of those phenomena seldom are different characterizations of the very same states of affairs. Individuals' racist views may causally influence the features of social systems, and similarly for interspecific competition and evolutionary trajectories, but in neither case do the former compose, realize, or determine the latter.

Debates about explanatory reduction have tended to conflate two questions, roughly: (1) how finely (i.e., at what "level") explanatory factors should be characterized, and (2) whether explanatory factors tend to be local, perhaps even components of the system exhibiting the phenomenon to be explained (i.e., components at "lower levels"). Scientific research has empirically shown (2) to be wrong: explanations cite the distant, the largescale, and the structural in order to shed light on a variety of phenomena. This outcome should lead us to shift our gaze outward rather than down, so to speak, when looking for candidate explanations. And that,

[7] I suspect most if not all scientific explanations include causal information, but what I say here is not intended to commit one to the view that all scientific explanations are causal explanations.

in turn, leads the first question of reductionism, (1) above, to seem rather beside the point. If entirely different factors are targeted in different candidate explanations, then the question of how finely to characterize a factor does not arise. This relates to the point I made in the previous section that, when we consider candidate explanations actually generated in scientific research, they turn out not to be clearly related by determination, realization, supervenience, or composition.

Second, the framing of levels also suggests candidate explanations are less numerous than they in fact are. Looking merely to components on lower organizational levels suggests we will have, at most, the number of explanations as there are levels; for instance, the social explanation, psychological explanation, neurological explanation, and genetic explanation. When we shift our gaze outward instead of down, as I suggested just above, shifting our antireductionist expectation from levels to varieties of influences, this opens the door to the recognition of a much wider variety of candidate explanations. Candidate explanations, it seems, may be as numerous as factors that significantly bear on the phenomenon—or even as numerous as various partially overlapping sets of these factors that may be targeted in different investigations. Such candidate explanations differ not just in what factors are cited but also (as a result) in what circumstances or to what varieties of phenomena they apply. This is anticipated by what I said in Section 2 about representations with incommensurate generality, i.e., that generalize across different ranges of circumstances. Indeed, this is hardly surprising when we take into account the different research projects within which different explanations are formulated. Positing that there are potential explanations at different levels does nothing to resolve which of *these* potential explanations are better (and in which circumstances). And, then, the need for resolution on this broader question renders the question of better level of explanation rather redundant.

For example, explaining some phenotypic trait, say, variation in coloration in Harris sparrows, can take place in the context of research into frequency-dependent selection, explaining this as an instance of the hawk-dove game dynamics (Maynard Smith, 1984). Or an explanation may be generated in research on phenotypic plasticity, with an explanation that bears on this as an instance of environmental influence on trait development. (See Potochnik, 2016, for a more extended discussion of this example.) There are many more possibilities beyond these two: explaining trait variation within a population is of interest in a number of biology research programs. Oftentimes the difference is associated with differently characterized explananda, but specification of explanandum isn't sufficient to single out just one—or even a few—explanations (Potochnik, 2016). And nothing is special about this example. Relevance across multiple research projects and variable significance for those research projects is common for phenomena scientists aim to explain.

Third, this wide variety of candidate explanations and how they relate also complicates the grounds for deciding among the candidates. Classically, for antireductionism couched in terms of levels of explanation, advantages like generality, breadth, and stability have been touted as grounds for preferring non-reductive explanations. (For some recent discussions, see Weslake, 2010; Blanchard et al., 2018; Bradley, 2020.) I have suggested that, if varieties of explanation do not come in levels and do not bear special relationships to one another, then candidate explanations cannot be straightforwardly ordered with regards to generality. For antireductionism without levels, different varieties of generality suit explanations to contribute to different research projects, with different aims. We should not expect an across-the-board ordering for any other measures that may be relevant to an explanation's quality. Similar considerations may still play a role in determining which candidate explanation(s) fits the bill. But, given the ease of generating potential explanations with different forms of generality (i.e., that generalize to different ranges of systems), it is possible or even likely that such considerations will vary with the requirements of different research projects. The question may not be which explanation is most general, stable, or offers greatest breadth or guidance but, rather, which explanation has these properties in the right combination and regarding the right features to be most valuable to a specific research project.

If this is so, then this suggests a form of explanatory pluralism. One single explanation may not win out against all other candidates, but rather multiple explanations may be developed across science, each of which best addresses some research needs but not others. This is different from an explanatory pluralism developed within the levels framework, such as by Jackson and Pettit (1992) and Sober (1999), as those views adopt the expectation of levels of explanation ordered by relative generality and embrace pluralism with regards to how much generality is desirable. Note that one might follow my urging to reframe antireductionism without a commitment to such an explanatory pluralism: one may hold that there is always a single, best explanation that is non-reductive in the ways I have outlined (perhaps a single, integrative explanation that draws from all relevant research projects). On the other hand, if one does accept explanatory pluralism of this form, then this opens up a significant role for scientists' interests and priorities in shaping the nature of scientific explanations—due to what they emphasize and what they sideline in their particular research projects. This is explored for varieties of explanation in cognitive science by Potochnik and Sanches de Oliveira (2020), who call this different "explanatory styles."

To summarize, a better antireductionism about explanation stems from the insights that different potential explanations regularly feature entirely different factors influencing the phenomenon, that these potential explanations vary in what they attend to and what they abstract from as they are developed in and contribute to different research programs, and that grounds for deciding among these potential explanations include considerations that may also vary with different research programs and perhaps even with something as basic as scientists' interests. The antireductionism comes in granting the legitimacy or even preferability of at least some of these potential explanations that do not feature local microphysical happenings. Explanatory pluralism results from additionally asserting that multiple of these potential explanations are warranted for a single phenomenon (as characterized in some explanandum).

Nowhere in these statements of explanatory antireductionism and explanatory pluralism is reference to levels needed, and, I propose, such reference would actually be a liability. An antireductionism based on levels fails to incorporate these features and is impoverished by their absence. Further, even if the points made in this section are somehow accommodated, the levels framing remains problematic for philosophical debates about scientific explanation. Such a framing can easily slide into the presumptions I argued against in Section 2. This framing is also associated with other problematic and unearned ideas. For one, antireductionism based on levels of explanation has been taken to suggest that explanandum and explanans should be on the same "level," i.e., regard similarly sized objects operating at similar timescales (e.g., Wimsatt, 1972; 2007). This may work as a defense against reductionism, but it also defines away the possibility of large-scale and structural causes—and, for that matter, the possibility of tiny entities sometimes wielding great explanatory power. An instance of this is individualism in social science, where behavior is expected to be fully explained by the properties of individuals; see Haslanger (2016) for an argument against individualism in favor of structural explanation, or explaining behavior with reference to systems in which individuals participate. The expectation that explanations should match the level of what is being explained is clearly wrong. As discussed above, scientific explanations regularly cite the distant, the large-scale, and the structural in order to shed light on a variety of phenomena, as with structural explanation in the social sciences. It is a further liability of the levels framework that it obscures this in order to counter the view that all explanations trace back to microphysical happenings.

4 A New Working Hypothesis about Scientific Explanation

Oppenheim and Putnam's (1958) stance and Fodor's (1974) rebuttal were both explicitly formulated as "working hypotheses" about how the relevance of levels will play out in science: the former reductive unity, the latter independence of levels of realization. Both working hypotheses, I submit, have been proven wrong by scientific advances. At this point, there is ample scientific evidence in favor of antireductionism about scientific explanations. With few if any exceptions, the so-called special sciences continue about their business, indifferent to any breakthroughs in microphysics, and the explanations they produce are not treated as provisional, awaiting vindication by reduction. But, just as importantly, there is also ample scientific evidence that explanations don't come in levels. Different fields and subfields that target the same phenomena focus on different factors that by and large bear no straightforward metaphysical relationship to one another, and large-scale and systemic factors can be key to explaining many phenomena.

Both previous working hypotheses—reductionism and levels of explanation—share starting presumptions about how our world operates that turn out to be wrong. Both of these philosophical positions presume that the key to explaining phenomena is their features, their immediate causes, and perhaps what composes them; we might call this "localism" about explanation. Thus, the choice in explanation has been framed as between on-site microphysical happenings or lumpier characterizations of those happenings. But localism about explanation is wrong. Phenomena are regularly determined by large-scale and distant factors, by structural and contextual factors, by systems in which they participate. Abundant scientific and philosophical research supports this claim, including complexity research such as the aforementioned dynamical systems theory, developmental systems theory, systems biology, and network theory. Recall from above that one of the entry points for the levels framework in philosophy was in Herbert Simon's work on complexity. Historically, at least in that tradition, positing non-reductive levels of explanation was a way to accommodate complexity. But since then, the antireductionist levels framework and complexity research have parted ways: the former has retained a commitment to localism about explanation, while the latter is predicated on its rejection. And rightly so, it seems to me. Localism about explanation is demonstrably false.

This inspires the new working hypothesis I propose about scientific explanation. In accordance with the view outlined in Section 3, I propose that prospective explanations are many and varied, often including some featuring large-scale, distant, and structural influences. The decision is not about how fine-grained our characterization of local factors should be but rather about which factors at what scales we should attend to. This may well have at least some objective determiners, but I suspect some of the determination will be left to what scientists and their audiences prioritize, intentionally or not, via the specific research projects scientists pursue. This characterization fits better than either reductionism or antireductionist levels with the variety of potential explanations encountered in scientific research and with how those candidate explanations relate to one another.

The two diagrams of levels featured in above figures have held remarkable sway over our field, so I have tried to offer a competing image in Figure 2.3. It is more mundane: a failure to draw lines to demarcate levels or arrows to demarcate metaphysical determination relationships just shows up as blank page. Aside from what this image does not include, the important features are (1) that the explanandum is grouped with multiple different sets of related phenomena, (2) that those groupings are associated with the identification of different explanatory factors, and (3) these different explanatory factors bear no special relationship to one another

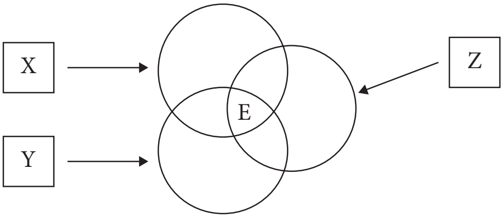

Figure 2.3 Antireductionism of explanation as a thesis about different explanatory factors operating at different scales.

(at least not in general). E is the explanandum; X, Y, and Z are factors that comprise candidate explanations for E.

Here is a brief example to illustrate this characterization. What explains the scarlet ibis’s bright coloration? (We could additionally specify the contrast class: rather than the white feathers of the closely related white ibis.) One candidate explanation focuses on the scarlet ibis’s ability to metabolize carotenoids. This highlights a primary form of coloration across bird species. Another candidate explanation focuses on the specific carotenoid carrier protein found in the scarlet ibis’s blood. Yet another candidate explanation postulates the role of the scarlet ibis’s vibrant coloration in mate attraction. Each of these is the subject of scientific research. Each distinguishes the scarlet ibis’s coloration from that of the white ibis. Each casts light on a different range of related phenomena: from avian coloration in general, to the scarlet ibis’s particular metabolism, to the role of bright coloration in sexual selection. See Figure 2.4. They are not competing explanations. One or another may turn out not to be exactly right, but it’s possible that all are correct. Depending on the specifics of our account of explanation, we may require more to be said about one or another for it to count as an explanation, favor one over others, or even anticipate their integration in a single explanation. However, I think the most likely outcome—and what best accommodates the realities of scientific investigation—is that all three of these explanations are accepted

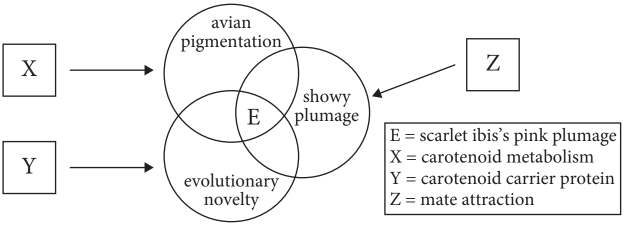
avian pigmentation
showy plumage
evolutionary novelty
E
X
Y
Z
E = scarlet ibis’s pink plumage
X = carotenoid metabolism
Y = carotenoid carrier protein
Z = mate attraction

Figure 2.4 Coloration of scarlet ibis illustration of different explanatory factors operating at different scales.

(or suitably refined versions if new evidence comes to light). But regardless, the point for present purposes is that the choice among potential explanations is not how finely to characterize the local details but rather which kinds of factors the investigation should target. Reductionism about explanation is incorrect—and so is antireductionism that relies upon levels, and the localism it presumes.

Acknowledgements

Thanks to Alastair Wilson and Katie Robertson for inviting my participation in this volume, as well as to Sahar Heydari Fard, Sally Haslanger, and the Structures Reading Group for helpful comments on an earlier draft.

References

Blanchard, T., Vasilyeva, N., and Lombrozo, T. (2018). "Stability, Breadth and Guidance." *Philosophical Studies* 175(9): 2263–2283.

Bradley, D. (2020). "Should Explanations Omit the Details?" *British Journal for Philosophy of Science* 71(3): 827–853.

Clarke, C. (2016) "The Explanatory Virtue of Abstracting away from Idiosyncratic and Messy Detail." *Philosophical Studies* 173(6): 1429–1449.

Craver, C. F. (2007). *Explaining the Brain: Mechanisms and the Mosaic Unity of Neuroscience*. London, UK: Oxford University Press.

Craver, C. F., and Bechtel, W. (2007). "Top-Down Causation Without Top-Down Causes." *Biology & Philosophy* 22: 547–563.

Eronen, M. (2015). "Levels of Organization: A Deflationary Account." *Biology & Philosophy* 30: 39–58.

Fodor, J. A. (1974). "Special Sciences; or, The Disunity of Science as a Working Hypothesis." *Synthese* 28(2): 97–115.

Franklin-Hall, L. R. (2016). "High-Level Explanation and the Interventionist's 'Variables Problem.'" *British Journal for the Philosophy of Science* 67(2): 553–577.

Garfinkel, A. (1981). *Forms of Explanation: Rethinking the Questions in Social Theory*. New Haven, CT: Yale University Press.

Gaston, Kevin J., and Blackburn, Tim M. (2000). *Pattern and Process in Macroecology*. Oxford: Blackwell Scientific.

Haslanger, S. (2016). "What is a (Social) Structural Explanation?" *Philosophical Studies* 173: 113–130.

Haug, M. C. (2011). "Abstraction and Explanatory Relevance; or, Why Do the Special Sciences Exist?" *Philosophy of Science* 78(5): 1143–1155.

Hempel, C. G., and Oppenheim, P. (1948.) "Studies in the Logic of Explanation." *Philosophy of Science* 15(2): 135–175.

Jackson, F., and Pettit, P. (1992). "In Defense of Explanatory Ecumenism." *Economics and Philosophy* 8(1): 1–21.

Kim, J. (1998). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. MIT Press.

Kim, J. (2002). "The Layered Model: Metaphysical Considerations." *Philosophical Explorations* 5: 2–20.

Longino, H. (2013). *Studying Human Behavior: How Scientists Investigate Aggression and Sexuality*. Chicago: University of Chicago Press.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge University Press.

Oppenheim, P., and Putnam, H. (1958). "The Unity of Science as a Working Hypothesis." In H. Feigl et al. (eds.), *Minnesota Studies in the Philosophy of Science*, vol. 2, Minneapolis: Minnesota University Press, pp. 3–36.

Potochnik, A. (2010). "Levels of Explanation Reconceived." *Philosophy of Science* 77: 59–72.

Potochnik, A. (2011). "A Neurathian Conception of the Unity of Science." *Erkenntnis* 34: 305–319.

Potochnik, A. (2016). "Scientific Explanation: Putting Communication First." *Philosophy of Science* 83: 721–732.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

Potochnik, A., and Sanches de Oliveira, G. (2020). "Patterns in Cognitive Phenomena and Pluralism of Explanatory Styles." *Topics in Cognitive Science* 12: 1306–1320.

Potochnik, A. (2021). "Our World Isn't Organized into Levels." In D. S. Brooks, J. DiFrisco, and W. C. Wimsatt (eds.), *Levels of Organization in the Biological Sciences*. MIT Press, pp. 61–76.

Putnam, H. (1975). *Philosophy and our Mental Life*. Cambridge: Cambridge University Press.

Simon, H. (1962). "The Architecture of Complexity." *Proceedings of the American Philosophical Society* 106: 467–482.

Sober, E. (1999). "The Multiple Realizability Argument against Reductionism." *Philosophy of Science* 66(4): 542–564.

Thomasson, A. (2014). "It's a Jumble Out There." *American Philosophical Quarterly* 51: 285–296.

Weslake, B. (2010). "Explanatory Depth." *Philosophy of Science* 77(2): 273–294.

Wilson, J. (2013). "Metaphysical Emergence: Weak and Strong." In Stephen Mumford and Matthew Tugby (eds.), *Metaphysics and Science*. Oxford University Press, pp. 345–402.

Wimsatt, W. (1972). "Complexity and Organization." In K. F. Schaffner and R. S. Cohen (eds.), *PSA: Proceedings of the Biennial meeting of the Philosophy of Science Association*. D. Reidel, pp. 67–86.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings*. Cambridge MA: Harvard University Press.

Yablo, S. (1992). "Mental Causation." *The Philosophical Review* 101(2): 245–280.

Ylikoski, P. (2014). "Rethinking Micro-Macro Relations." In J. Zahle and F. Collin (eds.), *Rethinking the Individualism-Holism Debate: Essays in the Philosophy of Social Science*. Springer, pp. 117–135.

3

How the Reductionist Should Respond to the Multiscale Argument, and What This Tells Us About Levels

Alexander Franklin

1 Introduction

Models and theories that describe interactions across many different spatial and temporal scales are ubiquitous in science. Such multiscale models are incompatible with a reductionist paradigm that assumes science can be neatly divided into distinct levels, and that sufficiently detailed understanding of dynamical models at lower levels will allow for the prediction of goings-on at all higher levels. This is the core of the multiscale argument against reduction.

In this chapter I respond to that argument by demonstrating that multiscale models undermine one form of reduction but are compatible with an alternate conception of localised reductive explanation, where smaller-scale details account for the explanatory adequacy of the multiscale models.

The multiscale argument and its variants is defended in Batterman (2013, 2020), Batterman and Green (2020), Bursten (2018), Jhun (2019), Massimi (2018), McGivern (2008), Mitchell (2009), and Wilson (2017) among others. Whereas the better known multiple realisability argument against reduction is usually premised on abstract metaphysical theorising,[1] multiscale arguments appeal to the details of science in practice. This research tradition does an excellent and important job of placing many often-overlooked aspects of scientific theorising under the lens of philosophy of science. I agree with many of the conclusions of these papers—it's only the implications for the reduction-emergence debate which, in my view, have been overstated.

Before proceeding, it's worth drawing a distinction between two kinds of antireductionist argument.[2] Arguments against methodological reductionism establish (in my view, successfully) that many different approaches to scientific reasoning are appropriate, legitimate, and successful, for different scientific projects; as such, methodological reductionists who might claim that top-down

[1] See Franklin (2021) for a discussion of how multiple realisation can be reductively ex plained.

[2] Robertson (n.d.) develops a similar distinction.

Alexander Franklin, *How the Reductionist Should Respond to the Multiscale Argument, and What This Tells Us About Levels*
In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Alexander Franklin 2024.
DOI: 10.1093/oso/9780192862945.003.0004

approaches are universally worse than bottom-up approaches are shown to be mistaken. However, many of the philosophers discussed below do not simply argue against methodological reductionism, they defend a form of worldly antireductionism: this is the view that more fundamental facts and relations are inadequate to explain the entities and regularities described by multiscale models.

I advocate methodological pluralism *and* a kind of non-eliminativist worldly reductionism: to evidence methodological pluralism I appeal to a successful multiscale model, where it's clear that a bottom-up methodology would fail; on the other hand, my evidence for worldly reductionism involves bottom-up explanations that, nonetheless, show why bottom-up modelling methodologies fail. Note that I call this reductionism 'worldly' rather than 'metaphysical' or 'ontological' to emphasise that it is compatible with the existence of non-fundamental entities.

The last bit of ground-clearing concerns claims about levels. Multiscale arguments effectively undermine the traditional hierarchical view of levels (see, e.g., Oppenheim and Putnam (1958)), since multiscale models straddle such levels. This poses a problem for traditional conceptions of levels that results in the following dilemma: either there are a great many more levels than often assumed, or level talk should be abandoned altogether. Opting for the second horn of this dilemma poses further questions for the reductionist concerning how to articulate reductive explanations.

To précis the positive argument of the chapter: reductive explanations explain the stability of various features at larger scales, in a piecemeal fashion from the bottom up. As such, they can tell us why it is that the multiscale models are successful, and why it is that methodological reductionism fails. Stability is explained, for each particular stable system or phenomenon, by detailing the structures and processes which legitimate the discarding of many degrees of freedom. That gives rise, in each case, to a descriptively accurate model which is stable with respect to perturbations in the discarded degrees of freedom. Importantly, such reductive explanations are far more localised than the theories that traditionally form the relata of reduction. That's why reductive explanations needn't make the false scaling assumptions that provide succour to antireductionists. Reductive explanations can evidence worldly reductionism and methodological non-reductionism: multiscale models describe the stable dependencies between distributed parts of large systems in virtue of which the higher-level explanations are available, but the stability of those dependencies can be explained from the bottom up!

Overall, the multiscale argument deserves more attention than it has received in the philosophical literature. As noted, my goals are not purely critical: I think that such science-in-practice analyses have a great deal to teach us about the nature of the world.

The critique of universal level stratification developed below has strong commonalities with Potochnik's chapter 2: I agree with her view that scientific explanations do not straightforwardly divide into levels, however I argue that a particular kind of reductive explanation is asymmetrically available, and that this is evidence for a kind of reductionism. There's also much in common with Knox's chapter's (Chapter 10) emphasis of the importance of abstraction in emergence and my claim that the robustness of variables is the primary *explanandum* of reductive explanation. A further consequence of the claims in this chapter—that if science can be divided into levels these will be local and contextually defined—is in agreement with the chapters by Bechtel (Chapter 7) and Kincaid (Chapter 9). In addition, part of the upshot of my reductionist approach is that reductive explanation can account for the robustness of the variables of non-fundamental sciences, and thus allows insight into the question of why there are non-fundamental sciences at all: as such, there are commonalities with the questions addressed in chapters by Bhogal (Chapter 16) and Strevens (Chapter 17).

In Section 2 I develop the multiscale arguments as found in the work of various philosophers. In Section 3 I develop a defence of worldly reductionism in terms of reductive explanation, and argue that this can avoid the problems with more traditional approaches to reduction. In Section 4 I cash this out with a case study taken from prominent multiscale arguments due to Batterman and Wilson; I go on to show that this case study is reducible on my account of reduction. In Section 5 I explore the upshots for philosophical accounts of levels. In Section 6 I conclude.

2 Multiscale Arguments

Reduction is often taken to imply that any theory which describes phenomena at a particular set of length-scales may be reduced to theories appropriate at smaller length-scales. Multiscale arguments purport to undermine such attempts at reduction. Such arguments work in two distinct ways, both of which will be discussed in this section.

First, reductionism is putatively undermined by ostension: it's demonstrated that some systems are so complicated, and straddle such a wide range of length-scales, that traditional reductionist approaches just have no hope of deriving their behaviour from the bottom up. Second, and more theoretically, multiscale arguments are used to bring out a range of invalid assumptions standardly employed in attempts at reduction. Together these arguments can demonstrate that reduction fails, and explain why it fails. To foreshadow the argument in Section 3: the explanation of why reduction fails opens the possibility that a more nuanced approach to reduction that does not employ such assumptions may succeed.

I focus on multiscale arguments against reduction due to Julia Bursten, Robert Batterman, and Mark Wilson. Each philosopher is naturalistically motivated and they all appeal to a combination of the two kinds of multiscale argument just mentioned. Insofar as their target is purely methodological reductionism, we have no grounds for disagreement.

I'll start by characterising multiscale models, and go on to say what the multiscale arguments are in more detail.

Let's start with a very simple multiscale model. Potochnik describes the relations between oak trees and squirrels. She notes that while oak trees are of the order of 100 times taller than squirrels, the study of population dispersal over time will describe their interactions: 'there is evidence that the rate at which oak populations spread is heavily dependent on the dispersal of seeds by squirrels' (Potochnik (2017, p. 184)). Meanwhile, for studies of population dynamics, individual trees interact with a population of squirrels: '[m]asting occurs when trees produce all their seeds in large bursts, which happens only in some years ... [o]ne squirrel does not eat enough seeds to drive trees to evolve masting; it takes an entire population' (ibid.). Thus, Potochnik's example demonstrates that if one is to model the ecology and dynamics of squirrel and oak populations accurately, a wide range of spatial and temporal scales are required.

Potochnik uses this example to make arguments about levels, to which I'll return in Section 5. However, this example suffices to get an idea of how a multiscale model undermines a simple reductionist view: if one were to attempt to derive the large-scale theory of oaks and squirrel populations from the small-scale theory of seeds and individual squirrels, one would go wrong because, it's supposed, masting and seed dispersal could not be predicted or explained. This account of reduction is clearly a caricature, and I wouldn't wish to attribute it to Potochnik. However, as we'll see in Section 3, it's not straightforward to formulate a more sophisticated reductionist framework to deal with such models.

Winsberg (2006, p. 142) defines multiscale as follows: '[t]he fact that [e.g.] three different theories at three different levels of description need to be employed makes the models "multiscale." The fact that these different regions interact simultaneously, that they are strongly coupled together, means that the models have to be "parallel multiscale"'. Note that Winsberg's 'level' should be read as 'scale'.

Winsberg goes on to explain that 'parallel multiscale' indicates that the different scales are interacting so strongly that they have to be modelled in parallel—if one aimed to express the input of one model in terms of a few parameters, and feed it into the other models, then such an approach would fail to be empirically adequate. We need to take into account the variation of details at all the relevant scales in a single model. However, Winsberg does not commit himself to a particular stance in the emergence-reduction debate. As such, while his account has informed the later literature, he does not make claims that I will challenge here.

On the other hand, Batterman (2013) (see also Batterman (2020)) specifically targets reductionism, which, he claims, is incompatible with accurate scientific modelling. Batterman stresses that one cannot understand the large scale *straightforwardly* in smaller-scale terms—thus he is in favour of methodological pluralism:

> Of course, the phenomenological parameters, like Young's modulus (related to Navier's ϵ), must encode details about the actual atomistic structure of elastic solids. But it is naive, indeed, to think that one can, in any straightforward way derive or deduce from atomic facts what are the phenomenological parameters required for continuum model of a given material. (Batterman (2013, p. 272))

Batterman draws an unfavourable comparison between a naïve scaling strategy—known as 'representative elementary volume' (REV)—and the renormalisation group (RG) strategy. He argues compellingly that, in certain contexts, the RG gets it right where the REV gets it wrong precisely because the RG employs a multiscale modelling strategy.

I don't wish here to discuss the RG in detail: see Franklin (2019, 2020) for discussion of this in condensed matter and quantum field theoretic contexts, respectively. Instead I focus on the more general objection to reduction raised in Batterman (2013).

This concerns the idealisations employed when attempting to do without multiscale models; the worry is that such idealisations often lead to inconsistencies between bottom-up and top-down approaches: in order to construct tractable models one generally assumes that a given system is homogeneous at scales smaller and larger than those of interest. Traditional approaches to reduction move from lower-level to higher-level descriptions by simple averaging or other techniques that build on the homogeneous idealisations and ignore the structures relevant at intermediate scales. Batterman argues that, in many cases, the use of such averaging techniques leads to inaccurate predictions which can be corrected only by paying attention to such intermediate scales—multiscale models are exactly those models that pay attention to goings-on at multiple scales.

> I suggest that much philosophical confusion about reduction, emergence, atomism, and antirealism follows from the absolute choice between bottom-up and top-down modeling that the tyranny of scales apparently forces upon us. As noted, recent work in homogenization theory is beginning to provide much more subtle descriptive and modeling strategies. This new work calls into question the stark dichotomy drawn by the 'do it in a completely bottom-up fashion' folks and those who insist that top-down methods are to be preferred. (Batterman (2013, p. 257))

Homogenisation is the process by which we move from accurate atomic models to accurate large-scale continuum models. Batterman notes that this theory 'involves appeal to various geometrical properties that appear at *microscales* intermediate between the atomic and the macro' (ibid., p. 258); Batterman refers to such intermediate scales as the 'mesoscale'. Thus, he argues that standard reduction, which solely appeals to bottom-up explanation, is inadequate here, and that intermediate multiscale models are required.

Batterman notes that 'scientists do not model the macroscale behaviors of materials using bottom-up techniques' (ibid., p. 257). In-principle claims are anathema to Batterman, and he concludes from his methodological observations that the multiscale methodology employed by practising scientists provides good reason to dismiss reductionism.

Bursten (2018) makes the antireductionist argument more forcefully. She presents a detailed case study about the propagation of nanoscale cracks, and draws a fairly strong set of conclusions. While she admits that her characterisation of reduction is somewhat crude, she supposes that the reductionist would favour the smallest scale quantum-mechanical model to the exclusion of the others. She then notes that this model

> does not have the conceptual resources to account for many of the features of interest of the simulated system. There are phenomena captured in the snapshot that quantum mechanics cannot resolve with its lens. Pressure waves, elastic strain, and thermal fluctuations in a solid are macroscopic, or occasionally mesoscopic, phenomena. Thermal fluctuations in particular simply cannot be tracked by quantum-mechanical descriptions of a system, and to deny their genuine reality, as this reductionist lens would, is to willfully ignore how materials really behave. (Bursten (2018, p. 162))

As noted above, I agree with methodological antireductionism—that accurate modelling of systems also requires consideration of multiple different non-fundamental spatiotemporal scales. But a subtler form of reductionism, which asserts that the adequacy of the non-fundamental models can be explained from the bottom up—and so builds in an explanatory asymmetry—is not thus refuted. In fact, while Bursten repudiates reductionism of any variety, she goes on to argue (p. 164) that '[i]t is both possible and, for the purposes of multi-scale modeling, necessary, to develop accounts of how different theories at different scales can be constructively combined to model material behavior.' As such, the more piecemeal approach to reduction advocated below may be countenanced.

I agree that understanding multiscale reasoning is an important task for philosophy of science. I argue, however, that such reasoning can be explained and understood in a way that justifies a particular reductionist attitude. Namely, that the working parts of such models are empirically adequate because of legitimate abstractions away from more fundamental goings-on; and that the legitimacy of such abstractions can be reductively explained. In particular, I claim that it is the details of the smaller-scale models which explain why, for example, a description in terms of pressure waves is successful, even if such a description is not available in terms of the entities at small scales.

Bursten and Batterman's detailed arguments demonstrate, primarily by example, that straightforwardly or naïvely scaling up from a given description at various particular length-scales will fail to take into account the complex intermediate structures. Thus, multiscale models which take into account such structure are necessary to good science.

Mark Wilson's work likewise engages with multiscale models, and at length draws out the complexity of such models and their incompatibility with standard reductionist approaches. However, what's especially relevant here is that Wilson does more than either Bursten or Batterman to provide a theoretical account of why reduction fails. His work can therefore be used to draw out and clarify the putative incompatibility between reductionism and multiscale models.

Wilson explicitly has Nagelian reduction in view. The state-of-the-art Nagelian approaches take all the scientific dependencies at some higher level, re-express them in the terminology of the lower level, and derive them from the lower-level theory, while allowing that the derivation may only recover an approximation; see, e.g., Butterfield (2011), Dizadji-Bahmani, Frigg, and Hartmann (2010), and Schaffner (2013).

Wilson observes that when reductionists attempt their bottom-up derivation they fail to take account of certain ways in which theories work. For Wilson, a theory is a theory façade, which is, in his inimitable prose, 'an uneven pile of pasteboard cutouts that ably masquerade, from selected angles, for an integral metropolis' (Wilson (2006, p. 356)); see Figure 3.1.

Theories involve a series of loosely and complexly connected methods and problems which suit different domains. For example, classical mechanics, which is further discussed in Wilson (2013), treats rigid bodies, point masses, and continua as its fundamental objects in different contexts. The concept of the theory façade is not just applicable to classical mechanics, it is widespread in science. If we

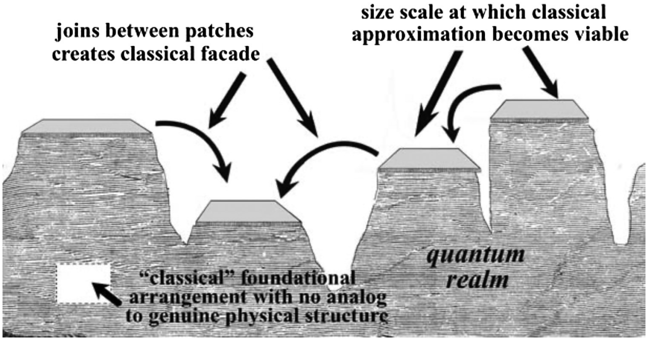
joins between patches
creates classical facade
size scale at which classical
approximation becomes viable
quantum
realm
"classical" foundational
arrangement with no analog
to genuine physical structure

Figure 3.1 A theory façade, from Wilson (2006, p. 196). This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

accept, following Wilson, that 'scientific theory' corresponds to a much more diverse, loosely bound notion than one might otherwise have assumed, the idea that one can reduce a theory is seen to be mistaken. At best one can reduce the localised application of different theories in particular contexts.

As such, care is required when importing the putative referents of theoretical terms from one context to another. For example, when creating a model of a particular system in a context, certain assumptions may apply: it may be permissible to model the system as if everything outside the region in question is uniform and that the environment can be captured by specific values for simple parameters. However, as Wilson emphasises, such simple parameters and assumptions often mask a huge amount of intricate detail. When redescribing the same system in a different context such assumptions will, in general, no longer apply, and the complexity of the environment may preclude description by simplified models.

For our present concerns: when attempting to model a given system from the bottom up, it is common to assume certain kinds of homogeneity—e.g., by simple averaging. Insofar as such assumptions are mistaken, bottom-up modelling goes wrong. Thus, by focusing on exactly how such assumptions can go wrong, Wilson urges a much more nuanced account of inter-scale relations in science.

As is shown in the case study in Section 4—discussed in Wilson (2017, Ch. 5), Batterman (2013), and Batterman and Green (2020)—assumptions that are justified in certain contexts idealise away intermediate scales and thus lead to false predictions in other contexts. Wilson aims to impress upon philosophers that modelling in science is a finicky process and that one should pay attention to the kinds of idealisation smuggled into each particular case. It's only very special systems that are well modelled as ideal gases or perfect crystal lattices for which we may accurately assume a homogeneous environment.

The arguments of Batterman, Bursten, and Wilson all raise doubts that attempts at reduction can adequately comprehend the nuance and complexity of real science. The correct reductionist response to these worries is to advocate a subtler approach to reduction: one that takes into account the intricacies of the mesoscale and is premised on localised reductive explanation.

3 Reductive Explanation

An important upshot of these criticisms of reduction is that science doesn't work in neatly stratified levels, and that expectations of the behaviour of some parameter in some context—e.g., that it's well represented at higher scales by simple averaging—will not in general carry over to other contexts. These, together, cast a shadow over the prospects for any grand reductionist project.

In this section I set out and develop an approach to reduction that is piecemeal, and very careful not to import warranted assumptions from one domain into another. The downside of piecemeal approaches to reduction is that one can no longer hope to complete the reductionist project and establish once and for all that reductionism is true or false.

What we can do instead is satisfy the more modest goal of increasing or decreasing credence in worldly reductionism by addressing those specific contexts in which antireductionist arguments have been put forwards. That is, where some have claimed that the properties of such and such are inexplicable from the bottom up, the more modest reductionist may step in with a bottom-up explanation for why such and such has those properties. The provision of bottom-up explanations is no trivial endeavour, and it's explicitly claimed by Bursten and Batterman that bottom-up explanations will often fail. As such, this is a meaningful form of reduction, and it's an important scientific and philosophical question whether or not it will succeed. Insofar as it does succeed, we will have established that a form of worldly reductionism is compatible with the multiscale complexity of the world.

In contrast to the Nagelian, I prioritise reductive explanation of (e.g.) the stability and autonomy of the non-fundamental models rather than wholesale derivation of their phenomena. This is for two reasons: first, in some contexts reductive explanation may increase one's credence in reductionism even while the science is insufficiently mathematised for derivation to go through; second, derivation sometimes requires the kind of unrealistic assumptions that the multiscale argument forces us to shun. Explanation, here, should be understood as appealing to a chain of worldly dependencies that relate the *explanans* to the *explanandum*; see, e.g., Woodward (2003).

With all that said, I'll explore this concept of reductive explanation in more detail, and show how this can account for the effectiveness of multiscale models in fairly abstract terms. In Section 4 I'll apply this reasoning to a multiscale model case study.

The central function of the reductive explanations considered here is to explain why the variables used to model dynamics in a given system are well suited to this job. Where a system is well described by a multiscale model, a number of variables, corresponding to different scales, will account for a system's interactions and dynamical evolution. Why these particular variables are the right ones to describe this system should be reductively explicable if some form of worldly reductionism can be evidenced. If it's not possible to explain why these variables work to describe this system—that is, if the success of a given multiscale model cannot be accounted for from the bottom up—then we have evidence against worldly reductionism. Any conclusions from such evidence will depend on the maturity and consequent warrant for realism of the relevant models and theories.

The way to explain why it is that certain variables are well suited for explanatory and predictive purposes is to identify processes and structures which pick out classes of variables as robust with respect to changes in underlying variables.

Take some set of variables $\{\mu_s, m_t, M_u\}$, found at micro, meso, and macroscales, respectively, that are used together in a multiscale model; the applicability of each variable will depend on its robustness with respect to various relevant perturbations; reductive explanation is achieved if it can be demonstrated, from the bottom up, which processes and structures are responsible for the robustness of each variable. This might involve, for example, the demonstration that M_u is robust with respect to perturbations in some subset of $\{\mu_{\{i\neq s\}}, m_{\{i\neq t\}}\}$ and that m_t is robust with respect to perturbations in some subset of $\{\mu_{\{i\neq s\}}\}$.

Multiscale models work by describing dynamics among a set of variables which is strictly smaller than the set of all variables that describe the system at all scales. Reductive explanation then tells us why it works to use the multiscale model rather than just having to describe the system in its entirety while keeping track of the interactions of every atom and every electron etc.

Note that, when referring to variables, I intend this plurally: variables feature in mathematised sciences, as well as in less formal scientific descriptions. Variables refer to worldly degrees of freedom, and, as such, the demonstration that a variable is stable with respect to perturbations is reason to believe that the associated degree of freedom has corresponding stability.

Note in addition that reductive explanations do not take any specific phenomena as their *explananda*. Rather reductive explanations explain the salience and goodness of the variables used to describe such phenomena. Therefore, if one has the relevant reductive explanations to hand, then this does not signal eliminativism. That's because reductive explanations do not explain the phenomena of the multiscale model; rather, they explain the effectiveness of that model for its target phenomena.

To recap: the multiscale argument raises problems with Nagelian reduction as traditionally applied. It shows that by treating a theory rather than localised applications of that theory, one idealises away crucial mesoscopic facts, and that interaction between different scales means that traditional reduction is impossible.

Reductive explanation has the potential to remedy these issues. The approach is far more localised than traditional Nagelian approaches as it accounts for the salience and effectiveness of particular variables for models in specific contexts—it thus bears commonalities with Rosaler (2017). In addition, it's explicitly targeted at explaining why multiscale models work—as such, it takes seriously multiscale interactions. The best way further to evidence these claims is by example, which I offer below.

Predictively accurate science can proceed in ignorance of the details of the composing materials or the goings-on at shorter temporal and spatial scales—many such details are irrelevant to multiscale model prediction and explanation. The reductive question is: can we explain such irrelevance of detail from the bottom up? Reduction is thus hostage to empirical fortune; reduction succeeds only in those circumstances where such explanations go through.

One advantage of the piecemeal approach to reduction advocated here is that it is clearly ontologically non-eliminativist. That's in part because the larger scale and multiscale models group and organise the world in a different way than their multiple reductive bases, but primarily because multiple reductive bases mean that no single lower-level description can supplant the multiscale description and its associated ontology.[3] Unlike with grander wholesale approaches to reduction, there is no single base theory that might purport to supplant the reduced theory.

4 Dislocations and Train Tracks

The aim of this section is to present a multiscale model to which traditional Nagelian reductions are ill suited but where reductive explanation sheds light—this establishes the compatibility of my claims that this model is amenable to reduction with the views of those philosophers who claim that it's not, assuming that they had a Nagelian (or similar) approach to reduction in mind. The case study of dislocations is especially apt for assessing the consequences of the multiscale argument against reduction—in a clear physics-based example it seems to show the use of top-down reasoning which cuts against reductionist bottom-up intuitions and demonstrates the issues faced by methodological reductionism. Thus it's worth investigating in detail.

The discussion in this section is largely drawn from Wilson (2017, Ch. 5), with the more detailed physics and maths drawn from Fan (2011), Friedel (1979), and Lu (2005). Figures 3.2 and 3.3 ought to be helpful in understanding the example.

Figure 3.2 depicts steel at various different length-scales. The scientific descriptions at these different length-scales interact in non-trivial ways. By paying attention to a variety of treatments at intersecting scales one can paint a picture that explains the relevant properties of steel bars or train tracks.

Wilson observes that, before the structure of dislocations was understood, the resistance of steel to breaking was mysterious. If one were to model steel simply, with a uniform lattice structure throughout, then one would predict that strains above a certain threshold would lead to large-scale steel deformation. That is, one would expect steel to be far more brittle than it in fact is. Wilson appeals to this case study in order to demonstrate the extent to which the assumptions of mesoscopic homogeneity may go wrong. It was only with the discovery of the mesoscale structure—the dislocations—that it was understood why steel is not brittle. Thus, only where the homogeneity assumptions are violated and, as such, the mesoscale structure is taken into account are accurate predictions for steel deformation available.

Dislocations are a general term for various ways in which an otherwise regular lattice is locally irregular; see Figure 3.3 for some examples of dislocations in two-dimensional crystal lattices. The general idea is that steel train tracks can be

[3] This idea is very closely related to the cross-classification discussed in Franklin and Robertson (2022).

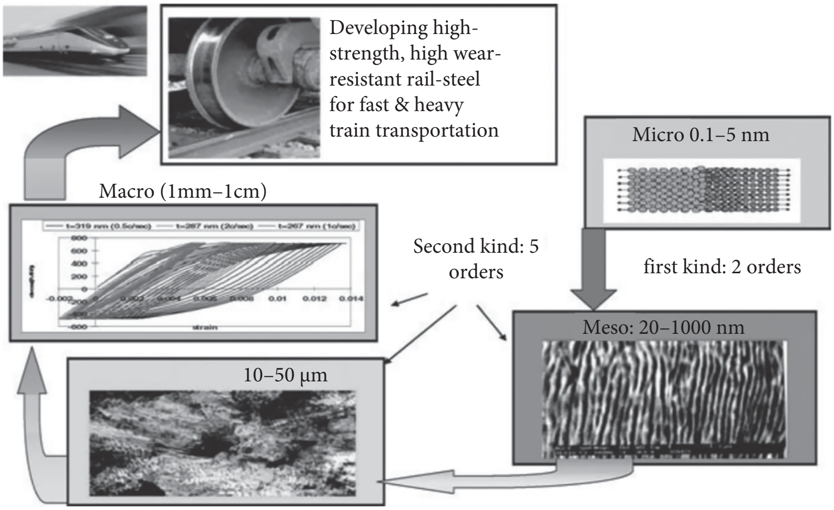

Figure 3.2 The multiple scales at which high-strength, high-resistance rail steel must be modelled. From Fan (2011, p.10) and Fan, Gao, and Zeng (2004). This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

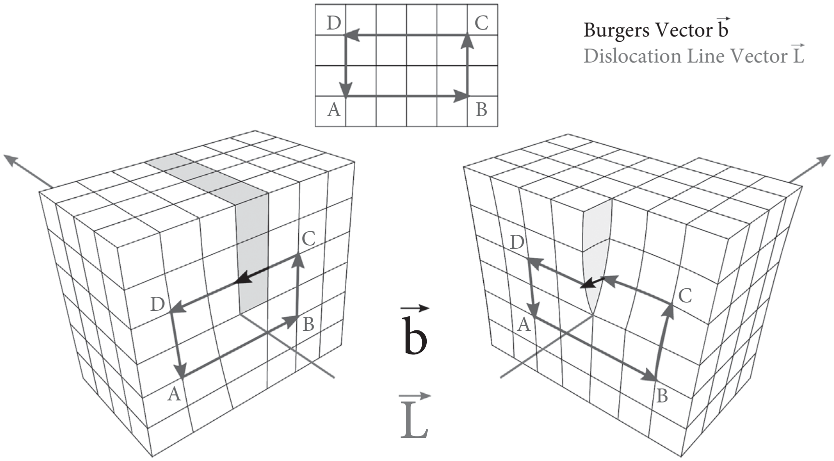

Figure 3.3 Dislocations. By Martin Fleck. Own work, distributed under the terms of the CC BY-SA 4.0 International licence, https://commons.wikimedia.org/w/index.php?curid=96858277.

manufactured with multiple dislocations throughout and that dislocations can move through the material. When the material is struck, or a force is otherwise imposed, the energy is dissipated via the motion of dislocations. While one might expect materials to change shape if struck, the ability approximately to retain shape depends on such motion.

> [The dislocations'] easy-to-achieve movements shield the underlying molecular bonds from the shearing distortions they would otherwise experience if the full impetus of the original blow had been allowed to reach their bonding sites directly. The net result is that RVE [representative volume element] units containing a plentitude of dislocations generally retain their dominant upper-scale behaviors far longer than they could if the dislocations weren't there, due to the fact that the dislocations significantly lessen the danger of fracture at the molecular lattice level. (Wilson (2017, p. 212))

Steel may have additional structures known as 'cementite'—10–50μm in Figure 3.2—through which dislocations cannot move; this explains why in certain circumstances steel may be brittle. The applicability of models used to describe the steel system is, thus, sensitive to the particular position of the dislocations. As long as most dislocations are relatively far from the cementite barrier, we may understand the dislocations as free, and therefore characterise the steel rail as sufficiently ductile to accept significant stresses. However, if the dislocations face a cementite barrier, a wholly different model with different kinds of assumptions is required to characterise the system's properties and responses to stress.

> [W]e cannot develop an adequate account of [steel] rail hysteresis working upwards from the molecular scale in a naïve manner [i.e. by simple averaging]. Multiscalar models evade these computational barriers by enforcing a cooperative division of descriptive labor amongst a hierarchy of RVE-centered submodels, each of which is asked to only worry about the dominant behaviors arising within its purview. (Wilson (2017, p. 221))

To summarise the case just considered: the scientific problem was that assuming mesoscopic homogeneity of steel led to inaccurate predictions for the hardness of steel; the resolution was to take into account mesoscopic structure—dislocations—which are responsible for absorbing energy from applied stress; this led to more accurate predictions of the fracture threshold of steel. The explanation of such improved predictions requires taking into account structure which is putatively left out in certain attempts at Nagelian reduction. Overall, the idealising assumption that microscopic symmetry carries on all the way up was shown to be mistaken.

This example clearly involves multiscale dependencies: the macroscale properties of the steel depend on the conditions at mesoscales and microscales. In order to determine the fragility of a given piece of steel accurately, one needs knowledge about its dislocation structure and the location of cementite barriers.

The challenge to traditional approaches to reduction arises from the tendency among its practitioners to make a large variety of homogeneity assumptions, which license scaling up the microscale description. In this context, such assumptions lead to inaccurate predictions.

4.1 Case Study Reduced

The case study shows that it's a mistake to assume that one can simply ask 'what will happen when this lattice undergoes this stress?' by scaling stress down and considering the effect on a small, regular segment of the lattice. To do so would hugely overstate the brittleness of the material. In other words, one needs to correct the scaled-down parameters by taking into account the intermediate structure of dislocations. One can't straightforwardly smear out the mesoscopic structure. Motivated by such considerations, from this and other case studies, Bursten and Batterman argue that reduction fails; this is based on the assumption that reduction requires a conception of the structure of materials which fails to take into account the true complexity of multiscale dependencies. In this section, I challenge that assumption.

It's important to note that Wilson is not antireductionist in the same sense as the other philosophers whose views I considered above. While he is similarly critical of those who discuss idealised theories—he claims they suffer from 'theory T syndrome'[4]—he is quite sympathetic to approaches which seek to evidence reduction by articulating the intricate interdependencies of various scientific models. In many respects, my approach advocated here is consistent with Wilson's observations.

In order to describe material structure adequately, it is often necessary to take account of communicating interconnected submodels: this will involve, for example, characterising the mesoscopic structure in terms of interlocking parts of cementite and dislocations. The question of interest in this section is whether one may explain, from the bottom up, the explanatory and predictive success of each interlocking mesoscopic model and of the cooperation of these models.

My claim is that one can show from the bottom up how the salient variables are robust and what leads to their playing roles in successful models. Although reductive explanation does not undermine many of the philosophical observations

[4] See Wilson (2021, Ch. 7) for more details.

raised in Section 2, it blocks the general objections to worldly reductionism. Since I do not seek to defend methodological or ontologically eliminativist forms of reduction, I don't need to show how we can describe each system entirely in terms of its smallest scale components; I rather aim to establish that each multiscale model variable is robust as a consequence of small-scale dependencies, and that there is no in-principle barrier to bottom-up derivation as long as that's sufficiently piecemeal.

So let's return to dislocations and consider how their description relates to that of the underlying lattice. I'll go on to ask whether or not reductive explanation goes through in this context.

Dislocations correspond to various types of localised disturbances to the lattice symmetry. They can be mobile and move through the lattice. They are often holes or gaps in the underlying lattice configuration. The relation between the dislocation variables and the underlying atomic lattice is illustrated in Figure 3.3.

While the full story is far more complex than I can discuss here, the central features on which I focus are that the dislocations are irregularities within the lattice and that they may travel through the material. These may be further understood by appealing to the Peierls-Nabarro (P-N) model; see Lu (2005) for an overview. This model allows for quantitative analysis of the size of dislocations and the force required for their motion. The model is a quasi-continuum model in that it selectively treats the lattice as composed of discrete atomic sites, and as a continuum. The continuum treatment, where used, allows for a considerably simpler model, although in certain places this leads to quantitative errors some of which are corrected in the Semi-Discrete P-N model. There are many ways to model dislocations; my goal here is to show one way that this is done and to discuss the extent to which this can be used to evidence a certain form of worldly reductionism.

In equilibrium, according to the P-N model, the distribution of atoms that constitutes a dislocation is determined by two distinct, competing contributions.[5] It costs energy to move atoms out of their positions in the equilibrium regular lattice. One part of this energy cost is the generalised stacking fault (GSF) energy: this is the sum of the misfit energy cost due to dislocated atoms and corresponds to a restorative force which attempts to make the dislocation smaller.

The elastic force opposes the restorative force and corresponds to the elastic energy. If one imagines the lattice split into two elastic half-spaces on either side of the dislocation, the movement out of regular equilibrium of all atoms on either side of the dislocation incurs an energy cost. Thus the elastic force attempts to make the lattices regular on either side of the dislocation, and as such to increase the size of the dislocation.

[5] Dislocations may move in two ways: either by gliding or by climbing. I will only consider dislocation motion in the glide plane. In addition, I only discuss translation dislocations.

The model describes these forces mathematically, minimises the total energy, and thus predicts an equilibrium structure of the dislocation, in particular the half-width of the dislocation core. Equation (1) describes the total dislocation energy on this model for dislocation density $\rho(x)$, generalised stacking fault energy $\gamma(\delta(x))$, elastic factor K, and long-distance cut-off L.

$$U_{tot}\left[\rho(x)\right]=\overbrace{\int_{-\infty}^{+\infty}\gamma\big(\delta(x)\big)dx}^{\text{misfit energy}}-\overbrace{\frac{K}{2}\int_{-L}^{L}\int_{-L}^{L}\rho(x)\rho(x')\ln|x-x'|dxdx'}^{\text{elastic energy}} \tag{1}$$

This model is fairly simple but predictively useful; it allows us to determine that the dislocation size depends only on fairly few smaller-scale details. The fact that dislocations are the result of a stable trade-off between two energetic factors underlies the efficacy of the dislocation model in describing mesoscopic structure.

The expression thus derived is useful for predicting dislocation size but doesn't tell us about dislocation motion. This is because, despite its derivation, it's a continuum model which is invariant with respect to spatial translations. One reason dislocations are interesting is because their motion can absorb energy; if there were no resistance to motion then that would not happen. The replacement of the misfit energy integral with a sum over the energy at each atomic position resolves this problem.[6] This takes us to a model where the dislocation moves through a series of potential wells. For a dislocation to move on this model it must overcome the Peierls barrier. The stress required to overcome this barrier is known as the Peierls stress (σ_p) and may be derived from this model; it is written as in equation (2) where μ, ν are elastic constants, d is the interlayer distance between the planes along which dislocations 'glide' (otherwise they 'climb'), and b is the Burgers vector (see Figure 3.3).

$$\sigma_p=\frac{2\mu}{1-\nu}\exp\left(-\frac{2\pi d}{b(1-\nu)}\right) \tag{2}$$

An interesting consequence of this equation is that the Peierls stress depends sensitively on d/b. For materials like ceramics which have low d/b, σ_p is too high for dislocation motion to prevent the material's fracturing. On the other hand, metallic systems have high d/b and thus are relatively ductile.

Having spelt out aspects of the derivation of the dislocation description, I turn to reductive explanation. I seek bottom-up explanations of the robustness of dislocation variables and their explanatory and predictive success. One upshot of the

[6] Note that the elastic energy still depends on a continuum assumption, though this can be corrected in the semi-discrete P-N model.

following analysis is that one can also use this bottom-up explanation to account for the limits of applicability of multiscale models that involve dislocations.

The dislocation description applies to a wide range of different underlying conditions. This can be seen by examining equation (1): the width of the dislocation core is derived by considering the competition of misfit and elastic energies. Determining this competition and minimising the total energy requires few details from the underlying system; as such the description is robust across a range of different values for other variables. However, importantly, the fact that dislocations are stable is determined by facts about the underlying lattice structure. Although the functions for elastic and misfit energy were expressed using continuous smaller-scale variables, their derivation explicitly requires and depends upon reasoning about properties of the atomic lattice.

While a more detailed account of the processes which lead to robustness would bolster this reduction, the details given here should be sufficient to undermine the multiscale argument against worldly reductionism—these establish that the properties of the mesoscale structure are understandable from the bottom up. Many of the possible motions of the underlying atoms are irrelevant to the dislocation description, and salient features of dislocations are a fairly straightforward consequence of the structure of the lattice. As such, the processes which lead to the lattice formation and the equilibrium atomic bonding ensure that many of the atomic displacement variables are irrelevant.

One particular function of the dislocation model is to explain the ductility or brittleness of certain materials. This in turn depends on calculating the force required to effect dislocation glide. Crucial to understanding from the bottom up is that resistance to motion depends on the discreteness of the atomic lattice. Thus, the dynamics of the dislocation model are also a consequence of features of the lattice. This explanation works for a fairly wide range of smaller scale conditions—the value for the Peierls stress is dominated by d/b; thus, details of dislocation motion are insensitive to other underlying details. In other words, dislocation glide can be described without reference to many of the smaller-scale details. This establishes the stability of the mesoscale description in terms of dislocations and their properties.

We can then consider the range of conditions of the underlying system over which dislocations are robust. It is required that the rest of the lattice is relatively well ordered: if too irregular then the dislocation will be indistinguishable from the movement of all the atoms around it. Lattices will be regular for a wide range of temperatures below the material's melting point.

The prospects for reductive explanation look good. We have a derivation of the robustness of the dislocation variables which then feed into the multiscale model. The processes which collude to make these variables effective for predicting the ductility of steel may be explained from the bottom up: they depend on details of the inter-atomic forces and the lattice structure.

We may reject eliminativism about dislocations because these are essential to a class of scientific explanations. Further reason to include dislocations in our ontology is that their dependencies are distinct from lower-level dependencies and screen off various atomic motions. See Franklin and Robertson (2022) for the argument that dislocations are, thus, emergent.

Multiscale methods are essential to accurate modelling of materials: for example, in the above analysis figures for elastic constants and generalised stacking fault energies may be empirical or may depend on modelling at other scales; moreover, the concepts of brittleness and stress are generally defined at larger scales. But the fact of the utility of such multiscale techniques ought not to preclude our asking, where appeal is made to details at larger scales, whether such appeal may be explained reductively. As the discussion in this section has shown, such questions may be addressed even in the context of multiscale models. Moreover, the account in this section helps establish methodological non-reductionism: it would be a mistake for those scientists interested in predicting the ductility of steel to start with the atomic lattice, assume mesoscopic homogeneity, and scale up precisely because there is non-trivial robust mesoscale structure. As a consequence of such structure, methodological reductionism would go wrong in this context.

Were we to have been satisfied purely with Nagelian reduction that employed various idealisations, Batterman, Bursten, and Wilson's worries would remain unanswered. They establish that there is a complex and relevant mesoscale structure which determines the macroscopic properties of many materials. By offering a reductive explanation I have shown, from the bottom up, how and why such mesoscale structures are stable.

The dislocation model is appropriately embedded within a much larger framework and, if one wants to discuss reduction in the larger framework, one needs to make sense of widespread multiscale modelling. As my aim is not to defend eliminativism, I do not think that the use of such modelling practices is overly worrisome to the reductionist; nonetheless reductionists ought to engage in the difficult task of attempting to go as far as possible with reductive projects. Once we have established the details of the dislocation picture, this then ought to be related all the way up to the description of stresses applied to the macroscale steel structures. Of course I haven't gone that far, nor is the dislocation model discussed here the most detailed available; however, I have demonstrated that, even in contexts of multiscale dependencies, we can make progress towards evidencing worldly reductionism by offering localised reductive explanations.

5 Levels

One feature of the discussion in previous sections is worth noting: reduction usually involves reference to levels, where affairs at higher levels are reduced to those at lower levels, but thus far I have studiously avoided such terminology. This observation alone calls into question a certain antireductionist argument—that reduction requires a levels hierarchy which is incompatible with multiscale models. Notwithstanding that levels aren't required for reduction, the question I pose (but don't settle!) in this section is whether any adequate conception of levels appropriate to a multiscale world can be recovered.

Note that while I haven't talked about levels, reduction does have a preferred direction: reductive explanations account for less fundamental details in terms of more fundamental details. Greater fundamentality, in the contexts discussed here, generally corresponds to smaller scales, but there are exceptions to this, most obviously in cosmological contexts. A full account of fundamentality would take more space than I have available, but I follow McKenzie (2019) in supposing that the fundamentality relation should be *a posteriori*.

Having said that reduction can proceed while avoiding levels talk, one might nonetheless think that levels play a useful role for philosophy of science, so we should consider whether any conception of levels can withstand the multiscale argument.

Levels, as used in science and philosophy of science, may be taken to have three salient features that are of interest for the present discussion. I'll first explain these and then argue that at least one of them must be sacrificed given the widespread scientific appeal to multiscale models.

First, levels contain the resources to explain and predict much of what goes on at that level. That is, goings-on at a level are commonly explained or predicted by facts or details at the same level—many intralevel explanations and predictions are available. Call this 'effective explanatory/predictive closure'. Any account of levels that lacks this feature would fail to correspond to standard scientific usage; one reason that, for example, physics and biology are often assumed to inhabit different levels is that one can proceed with predictions and explanations in biology in relative ignorance of theories of physics, and vice versa.

Second, levels are linked to a fairly narrow range of spatial and temporal scales: examples of levels offered in the literature assume some limits in range such that one can change level by zooming in and out.

Third, levels uniquely partition the world such that it's determinate—and not context dependent—whether any two entities or facts share a level. Historically Oppenheim and Putnam (1958) endorsed this global account of levels. While more recent accounts focused on biological sciences have accepted a more contextual and local account—see references in Eronen and Brooks (2018)—the metaphysics and physics literature insufficiently emphasises this point. For example, while List's chapter in this volume is in principle compatible with a contextual and local characterisation of levels, this isn't explicitly mentioned.

The fact that many of our best scientific models are multiscale generates a conflict between the first feature on the one hand, and both the second and third features on the other. That's because multiscale models are required for a great many predictions and explanations; so granting the first feature implies that some levels are multiscale. But multiscale levels will be spread over a wide range of spatial and temporal scales, and different multiscale models would disagree on which entities share levels.

It's reasonable to suppose that accurate models of all worldly phenomena would include some multiscale models that link every scale to almost every other scale either directly or through a series of inter-related multiscale models. One then faces a dilemma: either we grant that levels are only defined locally or relative to a particular dynamical context, or we attempt uniquely to partition all entities with the consequence that the level including all inter-related multiscale models will cover an arbitrarily broad range of scales.

The case study in Section 4, and, in particular, Figure 3.2, provides a nice illustration of these issues. Predicting and explaining the brittleness of steel requires an understanding of the properties and dynamics of entities from the nanometre to the centimetre ranges. If these were all to be at a single level, this would violate the narrow range of scales feature of levels. It would also be at odds with the assumption that, in other contexts, one can explain the electrical conductivity of steel by focusing on a single level, corresponding to the structure at a much narrower range of scales.

The upshot of taking multiscale models seriously is that if there are any levels these should only be defined locally or contextually—relative to the dynamics of interest required to account for a given phenomenon. This account of levels will certainly allow for the recovery of some of the standard usage of the concept in science, but the cost may be that levels are so ubiquitous that the concept may seem fairly empty to many. Alternatively, if one were to insist on a more restricted and unique levels partition and did not want to give up on effective predictive/explanatory closure, one would likely end up with a single level covering the entire range of scales.

It's my view that aspects of this question are terminological—although I take questions of reductionism to be metaphysically substantive, whether one uses 'levels' in one way or another is less important. If levels were required for the reductionist project, then I would take these problems to be rather serious. However, it seems to me that we do very well in the philosophy of science without an account of levels that neatly fits all the many uses of the term.

6 Conclusion

Accurate scientific descriptions of the world involve a great deal of complexity. And science grows ever more complex with ever more caveats to the applicability of its models. Such trends are part of the motivation for the philosophers considered in this chapter. If, rather than growing more unified, the complexity of science is increasing, how could reductionism be maintained?

That question is behind the multiscale arguments considered above. I think that such arguments deserve more attention than they have received in the literature. And I've claimed that those arguments establish, firstly, that methodological reductionism is false, and, secondly, that the conception of levels often in use in the philosophical literature requires rethinking.

However, I also claimed that the view of reductionism assumed by defenders of the multiscale argument tends to be overly simplistic. Aside from methodological reductionism, they focus on a view that makes unwarranted idealisations. For example, they emphasise that putative reductions are predicated on homogeneity assumptions which, in many circumstances, drastically misrepresent the target systems. When solids are out of equilibrium or there is mesoscopic structure, we need to take a more nuanced approach to describing the world. I have argued that such nuanced approaches are consistent with worldly reductionism, evidenced by localised reductive explanations.

As a consequence, I demonstrated that the increasing complexity of science is compatible with reductionism so understood. Nonetheless my argument does not establish worldly reductionism—it's still an empirical question whether or not this should be accepted. Scientific investigation is required to establish if reductive explanations are available in every context.

Acknowledgements

Thanks to Al Wilson and Katie Robertson for the invitation to contribute to this volume, and to Eleanor Knox, Katie Robertson, Mark Wilson, anonymous referees, and audiences of graduate student seminars at King's College London and the University of Pittsburgh for feedback and comments on various versions of this paper.

References

Batterman, Robert W. (2013). "The Tyranny of Scales". In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press, pp. 256–286.

Batterman, Robert W. (2020). "Multiscale Modeling: Explanation and Emergence". In: *Methodological Prospects for Scientific Research*. Ed. by Wenceslao J. Gonzalez. Springer, pp. 53–65.

Batterman, Robert W., and Sara Green (2021). "Steel and Bone: Mesoscale Modeling and Middle-out Strategies in Physics and Biology". In: *Synthese*. 199, pp. 1159–1184. URL: http://philsci-archive.pitt.edu/17388/. Forthcoming in Special Issue on Multi-Scale Modeling and Active Materials, ed. Patrick McGivern.

Bursten, Julia R (2018). "Conceptual Strategies and Inter-theory Relations: The Case of Nanoscale Cracks". Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 62, pp. 158–165. DOI: 10.1016/j.shpsb.2017.09.001.

Butterfield, Jeremy (June 2011). "Emergence, Reduction and Supervenience: A Varied Landscape". *Foundations of Physics* 41.6, pp. 920–959. DOI: 10.1007/s10701-011-9549-0.

Dizadji-Bahmani, Foad, Roman Frigg, and Stephan Hartmann (2010). "Who's Afraid of Nagelian Reduction?" *Erkenntnis* 73.3, pp. 393–412.

Eronen, Markus I., and Daniel Stephen Brooks (2018). "Levels of Organization in Biology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Fan, Jinghong (2011). *Multiscale Analysis of Deformation and Failure of Materials*. Vol. 5. John Wiley & Sons.

Fan, Jinghong, Zhihui Gao, and Xiangguo Zeng (2004). "Cyclic Plasticity across Micro/Meso/Macroscopic Scales". Proceedings: Mathematical, Physical and Engineering Sciences 460.2045, pp. 1477–1503. Proceedings of the Royal Society of London. DOI: 10. 2307/4143156.

Franklin, Alexander (2019). "Universality Reduced". *Philosophy of Science* 86.5. DOI: 10.1086/705473.

Franklin, Alexander (2020). "Whence the Effectiveness of Effective Field Theories?" *British Journal for the Philosophy of Science* 71.4, axy050. DOI: 10.1093/bjps/axy050.

Franklin, Alexander (2021). "Can Multiple Realisation Be Explained?" *Philosophy* 96.1, pp. 27–48. DOI: 10.1017/S0031819120000285.

Franklin, Alexander, and Katie Robertson (2022). *Emerging into the Rainforest: Emergence and Special Science Ontology*. https://philsci-archive.pitt.edu/19912/.

Friedel, J. (1979). "Dislocations—An Introduction". In: *Dislocations in Solids: The Elastic Theory (Vol. 1)*. Ed. by F. R. N. Nabarro. North-Holland Publishing Company, pp. 1–32.

Jhun, Jennifer (2021). "Economics, Equilibrium Methods, and Multi-Scale Modeling". *Erkenntnis* 86.2, pp. 457–472.

Lu, Gang (2005). "The Peierls—Nabarro Model of Dislocations: A Venerable Theory and its Current Development". In: *Handbook of Materials Modeling*. Ed. Sidney Yip. Springer, pp. 793–811.

Massimi, Michela (2018). "Three Problems about Multi-Scale Modelling in Cosmology". *Studies in History and Philosophy of Modern Physics* 64, pp. 26–38. DOI: https://doi.org/10.1016/j.shpsb.2018.04.002.

McGivern, Patrick (2008). "Reductive Levels and Multi-Scale Structure". *Synthese* 165.1, pp. 53–75. DOI: 10.1007/s11229-007-9232-3.

McKenzie, Kerry (2019). "Fundamentality". In: *The Routledge Handbook of Emergence*. Ed. by Sophie Gibb, Robin Findlay Hendry, and Tom Lancaster. Routledge Handbooks in Philosophy. Taylor & Francis. Ch. 3, pp. 54–64.

Mitchell, Sandra D. (2009). *Unsimple Truths: Science, Complexity, and Policy*. The University of Chicago Press.

Oppenheim, Paul, and Hilary Putnam (1958). "Unity of Science as a Working Hypothesis". *Minnesota Studies in the Philosophy of Science* 2, pp. 3–36.

Potochnik, Angela (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Robertson, Katie (n.d.). *Autonomy generalised; or, Why doesn't physics matter more?* Forthcoming in Ergo. https://philsci-archive.pitt.edu/19911/

Rosaler, Joshua (2017). "Reduction as an *A Posteriori* Relation". *British Journal for the Philosophy of Science Volume 70, Number 1*. DOI: 10.1093/bjps /axx026.

Schaffner, Kenneth F. (2013). "Ernest Nagel and Reduction". *Journal of Philosophy* 109.8/9, pp. 534–565.

Wilson, Mark (2006). *Wandering Significance: An Essay on Conceptual Behaviour*. Oxford University Press.

Wilson, Mark (2013). "What Is "Classical Mechanics" Anyway?" In: *The Oxford Handbook of Philosophy of Physics*. Ed. by Robert W. Batterman. Oxford University Press. Ch. 2, pp. 43–106.

Wilson, Mark (2017). *Physics Avoidance: Essays in Conceptual Strategy*. Oxford University Press.

Wilson, Mark (2021). *Imitation of Rigor: An Alternative History of Analytic Philosophy*. Oxford University Press.

Winsberg, Eric (2006). "Handshaking Your Way to the Top: Simulation at the Nanoscale". *Philosophy of Science* 73, pp. 582–594.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford Studies in Philosophy of Science. Oxford University Press.

PART II

LEVELS OF EXPLANATION IN CAUSAL MODELLING

4
Exclusion Excluded

Brad Weslake

1 Introduction

I take the exclusion problem to be the problem of providing a principled reason to reject at least one of the following inconsistent claims[1]:

- **Non-Reductionism**. Mental properties are distinct from, though metaphysically necessitated by, physical properties.
- **Completeness.** Every event has a complete causal explanation in terms of physical properties.
- **Mental Causation**. There exist causal explanations of events in terms of mental properties.
- **Exclusion.** If an event has a complete causal explanation in terms of one set of properties, then it has no causal explanation in terms of any other properties.

In this paper, I examine the prospects for a principled rejection of **Exclusion**. Following Horgan (1997, 166) and Bennett (2003, 473; 2008, 283), I will refer to this position as *compatibilism*.[2] I will refer to the conjunction of **Non-Reductionism**, **Completeness**, and **Mental Causation** as *non-reductive physicalism*.

Compatibilism is a popular position.[3] However, it has frequently been defended in the absence of an independently justified general framework for thinking about causation and causal explanation. That began to change after the development of a justly influential theory of causation and explanation by James Woodward

[1] While this initial formulation of the problem involves the causal explanation of events in terms of properties, everything I say below could be reformulated depending on your preferred view of the causal relata. Sometimes **Completeness** is weakened, so that it does not presuppose that all events have complete causal explanations. I employ the stronger principle for simplicity, as it will not make any difference to my argument. I assume throughout that 'explanation' is a factive term, and that in a causal explanation all explanans properties are causes of the explanandum. If you prefer not to formulate the problem as involving explanation at all, but rather as involving complete or sufficient causes, be patient: explanation will not appear in the final formulation of the problem I reach in Section 3.5.

[2] Bennett restricts the term to those who say that mental causation is possible without causal overdetermination. I use the term in Horgan's more general sense.

[3] Bennett (2003) cites Goldman (1969), Blackburn (1991), Pereboom and Kornblith (1991), Yablo (1992), Burge (1993), Mellor (1995, 103–104), Horgan (1997), Noordhof (1997), and Yablo (1997), to take just a few of the more prominent adherents.

Brad Weslake, *Exclusion Excluded* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson,
Oxford University Press. © Brad Weslake 2024. DOI: 10.1093/oso/9780192862945.003.0005

(2003), which has come to be referred to as *interventionism*. The development of interventionism generated a robust debate concerning whether an interventionist is entitled to reject **Exclusion**, and it is this question I explore in what follows.[4] My central claim is that there is a significant blind spot in the existing discussion, concerning the nature of the relationship between physical and mental properties. Attention to this blind spot reveals that while the best formulation of the interventionist theory of causation entails the falsity of the exclusion principle, it does so at the cost of revealing a weakness in the interventionist theory itself.

The structure of the paper is as follows. In Section 2 I introduce interventionism. In Section 3 I consider how to formulate the exclusion problem in interventionist terms, addressing each component of the problem in turn. In Section 4 I turn to arguments for **Exclusion**. In Section 4.1 I introduce a principle, *subvenience sufficiency*, concerning the relationship between physical and mental properties. The existing discussion has universally accepted the principle, thereby accepting a position I call *internalism*. I consider exclusion arguments from that standpoint in Section 4.2. In Section 4.3 I formulate an exclusion argument under the assumption that subvenience sufficiency is false, a position I call *externalism*. I argue that while interventionism has a response to the argument, it is one that reveals a limitation in the interventionist theory itself. I conclude in Section 5.

2 Interventionism Introduced

Central to the interventionist framework is the notion of a causal model.[5] A causal model is a representational device for encoding counterfactual relationships between variables. Counterfactual relationships are represented by equations which specify the way in which the value of a single variable on the left-hand side would change as a function of changes to the values of the variables on the right-hand side. More formally then, a causal model is an ordered pair $\langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is a set of variables and $\mathcal{E}$ a set of equations, and every variable appears on the left-hand side of exactly one equation.

For example, a model $\mathcal{M}_1$ might contain equations representing that variable Y depends on variables X_2 and X_3, that variable X_2 depends on variable X_1, and that variables X_1 and X_3 took values 1 and 0 respectively:

$$Y := X_2 \vee X_3$$

[4] For arguments broadly sympathetic to interventionism on this score, see Shapiro and Sober (2007), Shapiro (2010), Raatikainen (2010), Polger, Shapiro, and Stern (2018), Woodward (2008a, 2015a, 2017), and Stern and Eva (2023). For arguments broadly critical, see Baumgartner (2010, 2009, 2013) (see also Baumgartner 2018), Hoffman-Kolss (2014), and Gebharter (2017b).

[5] Here I present just enough to set up the discussion that follows. For more extended introductions to causal models, see Hitchcock (2009, 2023).

$$X_2 := X_1$$

$$X_1 := 1$$

$$X_3 := 0$$

Here '$\vee$' should be interpreted as a function returning 1 if either side is 1 and 0 otherwise. Equations such as the last two, which simply assign a specific actual value to a variable, are *exogenous*. Equations such as the first two, which assign values as a function of other variables, are *endogenous*. I will assume that the equations are all deterministic, in which case the equations for a model entail the actual values of all variables in the model. In $\mathcal{M}_1$ for example, the equations entail that X_2 and Y both took value 1.

Variables in a causal model must represent entities capable of being changed by interventions, but the framework is otherwise consistent with a range of different metaphysical views concerning the nature of the causal relata. For simplicity, I will sometimes say that variable values represent properties and sometimes say that they represent events. All of this could be translated into whatever view of the causal relata is correct.[6] I will refer to a possible assignment of values to a set of variables as a *state* of that set, and I will talk freely of actual and possible variable values, changes to variable values, states, and changes of state of models. I will also talk about causal relations obtaining between variables and values of variables. This sort of talk should be interpreted throughout as reflecting corresponding actual or possible changes in, and causal relations obtaining between, what is represented by the model. I will assume throughout that a causal model must be veridical, in the sense that every counterfactual relationship specified by the model is true.

The counterfactuals represented by causal models concern *interventions*. An intervention is an exogenous change to the value of a variable in a model, in the sense that the values of the other variables in the model are not themselves causes or effects of the change, unless they are effects of the variable intervened on. Moreover, it is required that interventions be *surgical*, in the sense that the usual causes of the variable in question are suspended, so that the value of the variable depends only on the intervention. I will consider the nature of interventions in more detail in Section 4.2.

In the literature on causation it has been common to distinguish between type-causal relations and token-causal relations. An analogous distinction can be made between causal relations between variables and causal relations between variable

[6] For the complexities that this simplification evades, see Schaffer (2016, Section 1) and Gallow (2022, Section 1).

values. While the terminology is slightly misleading, I will follow Woodward (2003) and refer to causal relations between variables as *type-level* causal relations, and between values of variables as *token-level* causal relations.[7]

In the remainder of this section I introduce the definitions in the interventionist framework that will be important for what follows.[8]

First we need the type-level notion of a *direct cause* (Woodward 2003, 55):

- **DC.** X is a *direct cause* of Y in model $\mathcal{M}$ *iff* there is a possible intervention on X that would change Y when all other variables in $\mathcal{M}$ besides X and Y are held fixed at some combination of values by interventions.[9]

It is a necessary and sufficient condition for X to be a direct cause of Y in $\mathcal{M}$ that X appear on the right hand side of the equation for Y in $\mathcal{M}$. So for example in model $\mathcal{M}_1$, X_1 is a direct cause of X_2, and X_2 and X_3 are direct causes of Y.

Second, we need the type-level notion of a *directed path* (ibid., 42). This can be defined in terms of the properties of graphs associated with causal models. A *directed graph* for $\mathcal{M}$ is an ordered pair $\langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is a set of vertices that correspond to the set of variables in $\mathcal{M}$ and $\mathcal{E}$ is a set of *directed edges* connecting these vertices, where there is a directed edge from vertex X to vertex Y *iff* X directly causes Y in $\mathcal{M}$. The definition is then:

- **P.** A sequence of variables $\{V_1 \ldots V_n\}$ is a *directed path* from V_1 to V_n in $\mathcal{M}$ *iff* for all $i(1 \leq i < n)$ there is a directed edge from V_i to V_{i+1} in the directed graph for $\mathcal{M}$.

From here on, *path* should be read as equivalent to *directed path*. A path is simply a sequence of direct causes, but the graph-theoretic definition is useful because paths in a model can be easily discerned by constructing a diagram with the same structure as the associated directed graph. When presenting diagrams of this sort, I will follow the usual convention of using circles to represent vertices (variables) and arrows to represent directed edges (direct causes). So for example, by inspecting the diagram for $\mathcal{M}_1$ in Figure 4.1, it is easy to see that X_1 is a direct cause of X_2, that X_2 and X_3 are direct causes of Y, and that there is a path from X_1 to Y, from X_2 to Y, and from X_3 to Y.

[7] For a discussion of the relationship between type-causal relations, token-causal relations, and causal relations between variables, see Hausman (2005).

[8] While I provide references to Woodward throughout, the precise formulations I give are sometimes simplified or expanded, and sometimes make use of definitions introduced in this paper. One important simplification is that I am setting aside the generalisation to the case of probabilistic causation, on which see Fenton-Glynn (2021).

[9] In interpreting the condition in this way I agree with Baumgartner (2009). Woodward (2015a) confirms that this was his intended interpretation.

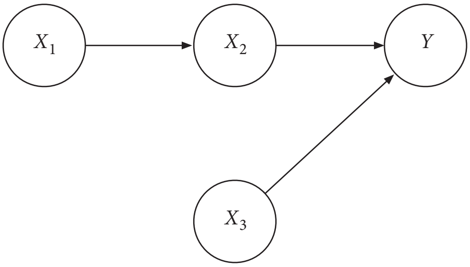

Figure 4.1 Diagram for $\mathcal{M}_1$.

I will provide diagrams of this sort when they are helpful. However, it is important to keep in mind that not all of the information relevant to causation in the interventionist framework can be read off these diagrams. In particular, to know whether the next two definitions are satisfied, you need to know the particular equations that relate the variables.

Third we need the type-level notion of a *contributing cause* (ibid., 59):

- **CC.** X is a *contributing cause* of Y in model $\mathcal{M}$ *iff* for some path P from X to Y in $\mathcal{M}$, there is an intervention on X that will change Y when all variables in $\mathcal{M}$ not on P are held fixed at some combination of values by interventions.

In model $\mathcal{M}_1$ for example, X_1, X_2, and X_3 are all contributing causes of Y. When $X_3 = 0$, an intervention setting X_1 from 0 to 1 would result in Y changing from 0 to 1. Likewise for X_2. And when $X_1 = 0$ and $X_2 = 0$, an intervention setting X_3 from 0 to 1 would result in Y changing from 0 to 1.

Finally, we need the token-level notion of an *actual cause*. The precise way to define actual causation in the interventionist framework remains a matter of lively debate. However, as I show in Weslake (unpublished), many of the proposed definitions can be formulated as instances of the following schema:

- **AC.** $X = x$ is an *actual cause* of $Y = y$ relative to model $\mathcal{M}$ *iff*:
  - **ACT.** The actual value of $X = x$ and the actual value of $Y = y$.
  - **PATH.** There exists a path P from X to Y in $\mathcal{M}$ for which an intervention on X would change the value of Y, when all variables $V_1 \ldots V_n$ in $\mathcal{M}$ that are not on P are held fixed at some combination of values satisfying <*conditions specifying permissible values* $v_1 \ldots v_n$ *for* $V_1 \ldots V_n$>.

The conditions specifying permissible values can be thought of as specifying the set of possible values of the off-path variables relative to which an intervention constitutes a test for actual causation along that path. All definitions of this form in the literature agree that *one* such permissible set is that in which all off-path variables have their actual values. So they all agree that a sufficient condition for $X = x$ to be an actual cause of $Y = y$ is for there to be a path from X to Y such that, holding all off-path variables fixed at their actual values, there is an intervention setting $X = x'$ where $x \neq x'$ that would result in $Y = y'$ where $y \neq y'$. In effect, that is, these theories agree that counterfactual dependence (of this sort) is sufficient for causation. Fortunately, for the purposes of the arguments I make below the differences between the various theories of actual causation on offer do not make any difference. So I will work with the following definition of actual causation, which also takes counterfactual dependence (of this sort) to be necessary for causation:[10]

- **AC**$_A$. $X = x$ is an *actual cause* of $Y = y$ relative to model $\mathcal{M}$ *iff*:
  - **ACT**. The actual value of $X = x$ and the actual value of $Y = y$.
  - **PATH**$_A$. There exists a path P from X to Y in $\mathcal{M}$ for which an intervention on X would change the value of Y, when all variables $V_1 \ldots V_n$ in $\mathcal{M}$ that are not on P are held fixed at their actual values.

In model $\mathcal{M}_1$ for example, $X_1 = 1$ and $X_2 = 1$ are actual causes of $Y = 1$, but $X_3 = 0$ is not. When we hold fixed $X_3 = 0$, an intervention setting X_1 from 1 to 0 would result in Y changing from 1 to 0. Likewise for X_2. But when we hold fixed $X_1 = 1$ and $X_2 = 1$, an intervention setting X_3 from 0 to 1 would not result in Y changing value from 1.

There are several consequences of these definitions that will be important in what follows. First, notice that if $X = x$ is an actual cause of $Y = y$ in $\mathcal{M}$, then X is a contributing cause of Y in $\mathcal{M}$. Second, notice that there may be more than one path that satisfies **PATH**$_A$. When **AC**$_A$ is satisfied in virtue of **PATH**$_A$ being satisfied by path P, I will say that $X = x$ is an actual cause of $Y = y$ *along path P*. Third, notice that each of these definitions is relativised to a causal model. The corresponding de-relativised definitions are as follows:[11]

- X is a *contributing cause* of Y *simpliciter iff* there exists a model in which X is a contributing cause of Y; and
- $X = x$ is an *actual cause* of $Y = y$ *simpliciter iff* there exists a model in which $X = x$ is an *actual cause* of $Y = y$.

I will return to the relationship between the relativised and de-relativised definitions in Section 4.3. However, because it will be important later, note that the

[10] The definition is equivalent (modulo some irrelevant differences) to Woodward's (AC) (2003, 77), and the definition of causation defined in terms of 'Act' in Hitchcock (2001, 286–287). In Weslake (unpublished), I argue against this and all other theories that fit the schema, but the arguments that follow also work for the theory I prefer.

[11] Here I follow Hitchcock (2007, 503) and Woodward (2008b). For an argument that interventionist definitions should not be de-relativised in this way, see Statham (2018).

de-relativised definitions do *not* require that for two variables to be causally related in either of these senses, *every* model containing those variables must represent them as causally related. One is enough.[12]

Because it will simplify the discussion later, I will also introduce the following de-relativised definition here:

- **Causal Chain**. There is a *causal chain* containing *X* and *Y* *iff* there exists a model in which *X* and *Y* are members of the same path.

This outline is sufficient to exhibit some of the key features of interventionism. First, the theory does not provide an analysis or reduction of causation but rather an explication of causal claims in terms of interventions. The concept of an intervention is itself clearly causal in character, and in the interventionist framework it is explicitly defined in causal terms. What is important for present purposes is that the truth of causal claims can be established independently of any such analysis or reduction—it is whether or not it is true that mental properties sometimes causally explain physical events that is at issue in the exclusion problem, not whether these explanations can be grounded in a reductionist account of causation. Second, this is a kind of counterfactual account of causation—causal claims involve what *would* happen given some particular intervention, not what *actually* or *will* happen. Third, causal claims are model-relative in the sense that they are only well defined with respect to the variables in a particular model. However, as should be clear, this is not a version of causal anti-realism. Causal claims are not made true or false by causal models, they are made true or false by the counterfactuals regarding experimental interventions that are represented by those models.[13] Moreover, because the counterfactuals are explicitly formulated in terms of interventions, it is typically transparent how they can be tested empirically. Nevertheless, as is clear from the definitions above, interventionism does entail that necessarily, if some causal claim is true, then there exists a model in which it is so represented.

3 Exclusion Reformulated

In the interventionist setting, the exclusion problem can be initially formulated as follows:

- **Non-Reductionism$_i$**. The values of mental variables are distinct from, though metaphysically necessitated by, the values of physical variables.

[12] In the terms employed by Stern and Eva (2023), interventionism so understood adopts the *Weak Causation Principle* but not the *Strict Causation Principle*.

[13] This may seem obvious, but the following mistake is routinely made (in this case, by a Nobel Prize winner): ‘A model is in the mind. As a consequence, causality is in the mind’ (Heckman 2005, 2).

- **Completeness$_i$.** For every event, there exists a causal model containing only physical variables which specifies a complete explanation of that event.
- **Mental Causation$_i$.** There exists a causal model in which a mental variable explains an event.
- **Exclusion$_i$.** If there exists a causal model specifying a complete explanation for an event, there exists no other causal model containing distinct variables specifying an explanation for that event.

In the remainder of this section I clarify and refine these notions in turn, and then more precisely reformulate the exclusion problem in the interventionist setting.

3.1 Non-Reductionism

Non-Reductionism$_i$ requires clarification of the distinction between mental and physical variables. It also requires clarification of the notion of a variable value being distinct from another variable value.

It is a standard presupposition in the debate over exclusion that there is a distinction between physical properties and mental properties in the sense required to generate the problem. I take no stand on how this distinction should be drawn. But to keep the relationship between models and what they represent clear, I assume that corresponding to it is a distinction between sets of variables that represent physical properties and sets of variables that represent mental properties. I will refer to these sets of variables as involving *vocabularies*, where a vocabulary is a set of variables with variable values that all represent a single property type. So a *physical vocabulary P* contains only variables with values representing physical properties, and a *mental vocabulary M* contains only variables representing mental properties.

In order for **Non-Reductionism$_i$** to occupy the proper role in the exclusion problem it needs to express a claim about the world, not about our ways of representing the world. So I will assume that variable values are distinct if and only if they represent distinct properties:

- **Value Distinctness.** Variable values are distinct *iff* they represent distinct properties.

In addition, note that there is a necessary condition on two variables appearing in the same model that follows from the definition of direct causation provided in Section 2. Recall that whether X is a direct cause of Y in $\mathcal{M}$ depends, according to **DC**, on whether there exists an intervention on X that will change Y when all other variables in $\mathcal{M}$ are held fixed at some combination of values by interventions. This implies an independence condition on variables coexisting in a model: if X is a direct cause of Y in $\mathcal{M}$, then there must be possible values x and x' for X such that an intervention on X from x to x' is possible when all other variables except Y in $\mathcal{M}$ are set to some combination of possible values by independent interventions. There is a natural generalisation of this independence condition standardly assumed to hold in causal models, which can be motivated by the idea that for any set of variables appearing together in a model it must be possible to non-trivially *test*, for every pair, whether **DC** holds. According to Woodward (2003, Section 3.5; 2015a), the relevant sense of possibility here is *at a minimum* metaphysical possibility. The corresponding independence condition on variables coexisting in a model $\mathcal{M}$ is this:

- **Independent Manipulability**. It is metaphysically possible that every proper subset of the variables in $\mathcal{M}$ be set to every combination of their possible values by independent interventions.[14]

Independent Manipulability reflects the natural idea that it is only variables not related by metaphysical necessity that are candidates for being related causally. It is well known that counterfactual theories of causation are inadequate if we allow dependencies between events that are related by metaphysical necessity (Kim 1973), and **Independent Manipulability** can be seen as the constraint that implements this restriction in the interventionist framework. When I refer to the interventionist theory of causation in what follows, I will take it to include all of the definitions provided in Section 2, as well as **Value Distinctness** and **Independent Manipulability.**[15]

My final formulation of **Non-Reductionism** is therefore:

- **Non-Reductionism**$_j$. Mental variables are distinct from physical variables in the sense that they are drawn from distinct vocabularies M and P, and the values of the M-variables are metaphysically necessitated by the values of the P-variables.

3.2 Completeness

Completeness$_i$ requires clarification of the notion of a complete explanation for an event. The exclusion problem is often framed in terms of causal sufficiency rather

[14] Baumgartner (2009, 167) calls a related condition *Fixability*, and Woodward (2015a, 316; 2017, 255) a related condition *Independent Fixability*. See Hoffmann-Kolss (this volume) for an additional line of argument for imposing the condition.

[15] For more detailed discussion of the reasons for imposing constraints of these sorts, and proposals for further necessary conditions on variables, see Hitchcock (2001, 2004, 2012), Halpern and Hitchcock (2010), Halpern (2016), Woodward (2016), Blanchard and Schaffer (2017), McDonald (forthcoming, 2023), and Hoffmann-Kolss (this volume).

than completeness, so any defensible notion of completeness must bear some close relationship to a notion of causal sufficiency. To begin, note that this should not be interpreted as being equivalent to the claim that every event can be completely explained in terms of some fundamental physical theory. This is so for at least two reasons. First, it is an open question whether reasonable candidates for fundamental physical theories should be interpreted causally.[16] Second, even if reasonable candidates for fundamental physical theories should be interpreted causally, it is not the case that the structure of fundamental physical theories is identical to the structure of causal models.[17] So for example, sufficiency is often defined along the following lines:

- **Physical Sufficiency**. An event *A* is physically sufficient for an event *B* *iff* the occurrence of *A* and the laws of physics together guarantee that *B* will occur (or fix a probability for *B* such that there are no further events conditioning on which would change the probability of *B*).[18]

However, **Physical Sufficiency** makes no reference to causal models, and it is not clear how it should be translated into those terms. Since I am proceeding under the interventionist assumption that causation is to be defined with respect to models, this notion is therefore inadequate for formulating the exclusion problem.[19] In making this point I do not mean to weaken the support that causal completeness assumptions rightly draw from the promise of complete explanations of events in terms of fundamental physical theories. My point is simply that there is an inference involved from the success of fundamental physics to the existence of a complete causal model in the sense required to formulate the exclusion problem in the interventionist setting.

Having clarified what **Completeness**$_i$ does not say, let us examine what it does say. There are a number of different notions of causal sufficiency that can be discriminated within the interventionist framework, only one of which, I will argue, is suitable for formulating the exclusion problem.

[16] See Russell (1913), Field (2003), and the essays in Price and Corry (2007).

[17] One reason is that causal models do not allow the representation of continuous processes (Strevens 2007, 242–244). Strevens puts the point by saying that interventionist causal models 'represent less of causal reality than is actually out there' (243), but an interventionist may consistently claim both that every interventionist model omits some causal truth, and that all causal truths are represented by some interventionist model or other (Woodward 2008b, 210–211).

[18] See, for example, Papineau (2001, 8; 2002, 17). Note that the relevant notion of event here must be liberal enough to allow events involving all physical properties instantiated across the entire cross-section of a light-cone in spacetime, if any events are going to turn out to be physically sufficient for any others (Field 2003; Ismael 2009, 2011).

[19] See Yates (this volume) for critical discussion of a set of principles closely related to **Physical Sufficiency**.

Consider first the following definition:

- **Sufficiency in the Circumstances in a Model.** A cause is sufficient in the circumstances for an effect in a model *iff* it is an actual cause of the effect in that model.

Since **Sufficiency in the Circumstances in a Model** collapses the notion of a sufficient cause and the notion of an actual cause, it is clearly too weak to play a role in formulating an appropriate causal closure principle. As a step in the right direction, consider next:

- **Weak Sufficiency in a Model**. Call an actual cause of an effect in a model a *weakly sufficient actual cause iff* it is an actual cause of the effect along path P, and there is no possible combination of interventions on variables not on P that would change the effect, if the actual cause were held fixed to its actual value by an intervention. A cause is weakly sufficient for an effect in a model *iff* it is a *weakly sufficient actual cause* of that effect in that model.[20]

Notice that this definition, like the preceding one, is a model-internal one, in the sense that sufficiency is defined with respect to a single causal model, and makes no reference to facts apart from those represented by that model.[21] This makes trouble. For suppose that we have a model $\mathcal{M}_P$ framed in variables drawn from physical vocabulary P which specifies a cause that is weakly sufficient for some effect. That is consistent with supposing that there exists some model $\mathcal{M}_{PM}$ constructed by adding variables from mental vocabulary M to $\mathcal{M}_P$, in which the M -variables specify an actual cause for the effect and the P-variables do not specify a cause that is weakly sufficient for the effect. What this possibility reveals is that the model-internal definitions of sufficiency do not adequately capture the idea, central to any closure principle, that when one class of properties is causally closed with respect to another class the latter do not make any *additional* causal difference. I conclude that an adequate closure principle must be at least as strong as:

- **Weak Sufficiency in a Weakly Closed Model**. Call a model $\mathcal{M}_F$ framed in variables drawn from vocabulary F *weakly closed* with respect to variables drawn from vocabulary G with respect to an effect *iff* $\mathcal{M}_F$ contains a weakly sufficient

[20] Related but distinct notions are *sustenance* in Pearl (2000, Section 10.2), *switch* in Woodward (2003, 96–97), *strongly causes* in Halpern and Pearl (2005, 855), and *sufficient condition* in McDermott (1995, 533; 2002, 96–97). To keep the formulations as simple as possible, I give definitions on which only values of single variables can be sufficient for others. The generalisation to multiple variables is obvious and does not make any difference to the arguments that follow.

[21] To re-emphasise a point I made in Section 2, remember that a notion being defined in a model-internal way does not imply that the corresponding fact is in any way model-dependent.

actual cause of the effect, and there exists no model $\mathcal{M}_{FG}$, constructed by adding variables from G to $\mathcal{M}_F$, in which any weakly sufficient causes in $\mathcal{M}_F$ are not also weakly sufficient causes in $\mathcal{M}_{FG}$. A cause is weakly sufficient for an effect in a weakly closed model $\mathcal{M}_F$ with respect to G *iff* it is weakly sufficient for the effect in $\mathcal{M}_F$.

Notice however that **Weak Sufficiency in a Weakly Closed Model** is compatible with the actual values of $\mathcal{M}_P$ specifying a weakly sufficient actual cause of an effect, and yet it being the case that no *alternative* values of the P-variables would have specified weakly sufficient causes of *alternative* values of the effect. That is, it is compatible with the actually instantiated physical properties sufficing for some event, while any alternatively instantiated physical properties would not have sufficed for any alternative event. So **Weak Sufficiency in a Weakly Closed Model** does not yet capture the sort of closure the successes of our scientific theorising typically license us to endorse, where for some class of properties proprietary to a theory, *whichever of those properties were instantiated* would have sufficed for all outcomes of a certain type. I conclude that the closure principle appropriate to formulating the exclusion problem in the interventionist setting is:

- **Strong Sufficiency in a Strongly Closed Model.** Call a cause *strongly sufficient* for an effect in a model *iff* it is weakly sufficient for the effect, and all alternative values of the cause would also be weakly sufficient for the value of the effect in any possible state of the model. Call a model $\mathcal{M}_F$ framed in variables drawn from vocabulary F *strongly closed* with respect to variables drawn from vocabulary G with respect to an effect *iff* $\mathcal{M}_F$ contains a strongly sufficient actual cause of the effect, and there exists no model $\mathcal{M}_{FG}$, constructed by adding variables from G to $\mathcal{M}_F$, in which any strongly sufficient causes in $\mathcal{M}_F$ are not also strongly sufficient causes in $\mathcal{M}_{FG}$. A cause is strongly sufficient for an effect in a strongly closed model $\mathcal{M}_F$ with respect to G *iff* it is strongly sufficient for the effect in $\mathcal{M}_F$.

It is important to note an immediate consequence of this definition. If a cause is strongly sufficient for an effect in a strongly closed model $\mathcal{M}_F$ with respect to G, then in any model $\mathcal{M}_{FG}$, constructed by adding variables from G to $\mathcal{M}_F$, there are no paths from any variables in G to the effect.

Strong Sufficiency in a Strongly Closed Model is a model-external definition but still a relative one, in the sense that a cause could be strongly sufficient in a strongly closed model with respect to one set of variables, but not with respect to a different set of variables.[22] While it would not make a difference to the argument below if we strengthened our understanding of completeness yet again, so that it

[22] This is also true of the closely related probabilistic conception of completeness in Sober (1999a, 139).

involved the idea of a model strongly closed with respect to *all* other variables, I prefer the present formulation. This is because understanding the problem in this way captures the great variability in the way completeness assumptions are formulated. Sometimes the worrying complete or sufficient explanation is supposed to be provided by physics, sometimes by biology, sometimes by neuroscience, sometimes by (at least the 'syntactical' explanations appearing within) cognitive science.[23] In my view the exclusion problem can be posed in terms of these different sciences precisely because it is reasonable to believe that there exist strongly sufficient causes, in models framed in variables drawn from the vocabularies of each of these sciences, which are strongly closed with respect to the M-variables.[24] If I had all of the physical information relevant to you, my knowledge of what you will and would do would not be increased by knowing any further mental information about you—and likewise if I had all of the biological information, or all of the neuroscientific information, or all of the ('syntactic') cognitive scientific information.[25] Moreover, once we understand completeness in the way I have suggested, it can be seen that the exclusion problem generalises—the physical causal model is strongly closed with respect to the variables of the biological causal model, the biological causal model is strongly closed with respect to the variables of the neuroscientific causal model, and so on up the hierarchy of the sciences and never vice versa.[26] And so if the exclusion problem arises for mental variables it also arises for any variables not appearing in some maximally strongly closed causal model.[27]

Because my central interest is in **Exclusion**, in what follows I will not defend these claims, and will simply proceed under the assumption that a closure principle concerning physical and mental variables formulated in terms of **Strong Sufficiency in a Strongly Closed Model** is true.[28] My final formulation of **Completeness** is therefore:

- **Completeness**$_j$. For every event, there exists a causal model with variables drawn from P, which is strongly closed with respect to M, in which there is a strongly sufficient actual cause P_1 for that event.

[23] The emphasis on physical causes is familiar from Kim (1998, 2005). As is made clear in Kim (1989), Kim was generalising from an argument initially formulated by Malcolm (1968) in terms of a 'neurophysiological theory'. The emphasis on syntactic causes is familiar from Field (1978) and Stich (1983).

[24] For a historical survey of how completeness in physics and biology became compelling, see Papineau (2001). For a comparison with the assumptions that generated earlier problems with mental causation, see Patterson (2005).

[25] See Loewer (2008, 2009). Note that this is not to say that my *explanations* would not be improved by the possession of this information. Indeed, I think they would be (Weslake 2010).

[26] It is a *hierarchy* in part *because* this relation is asymmetric in this way.

[27] Here I side with Bontly (2002) against Kim (1997, 1998).

[28] There are two options available to someone who accepts **Non-Reductionism**$_j$ but wishes to deny **Completeness**$_j$. One is to deny **Physical Sufficiency**, and with it **Completeness**$_j$. In my view the

3.3 Mental Causation

My initial formulation of **Mental Causation** is straightforward:

- **Mental Causation$_j$.** There exists a causal model with variables drawn from M in which a mental variable M_1 is an actual cause of an event.

Three comments on this formulation, before I introduce a revision in the following section.

First, interventionism is attractive not only as a theory of causation generally, but as a theory of mental causation specifically. In particular, it has been defended as providing a good framework for understanding causal explanation in psychology (Campbell 2007; see also Rescorla 2018; Kaiserman 2020), in psychiatry (Campbell 2008; Kendler and Campbell 2009), and in folk psychology (Menzies 2010). If interventionism is true, there is no special problem in understanding how a mental variable can be a cause.

Second, **Mental Causation$_j$** might be granted, and yet it might be argued that only a model containing physical variables *really* represents causes, and that models containing mental variables merely specify *explanations*, or some other weak cousin of causation. This might be because only the physical model promises to be maximally predictively accurate and therefore maximally strongly closed (Davidson 1963, 1967, 1970, 1995), or because the physical model is a truth-maker for the mental model (Crane 2008; Robb, Heil, and Gibb 2023, Section 5.3), or for more recherché metaphysical reasons (Jackson and Pettit 1988, 1990a, 1990b). In my view the arguments for these claims are unsound, but for present purposes I simply note that if they succeed they are arguments against interventionism in general and therefore should be addressed as independent claims about the nature of causation and causal explanation. Proceeding under the presumption of the truth of interventionism, I here set them to one side.[29]

Third, while this will also make no difference to the argument below, note that **Mental Causation$_j$** does not require that in order for mental variables to causally

most interesting arguments of this sort are those made by Cartwright (1983, 1994) as developed in the case of chemistry by Hendry (2006, 2010b, 2010a, 2017). See Sklar (2003) for the general line of response that I think blocks these arguments. The other is to accept **Physical Sufficiency** but deny that it entails **Completeness$_j$**. One strategy here turns on the idea that causes must be 'proportional' to their effects (Yablo 1992; List and Menzies 2009; Menzies and List 2010; Raatikainen 2010; and for critical discussion Weslake 2017). Another strategy, the 'dual explanandum' or 'intralevelist' solution to the exclusion problem, turns on the idea that the effects of mental causes are individuated mentally rather than physically (the position dates at least to Putnam 1975; see also Marras 1998; Thomasson 1998; Gibbons 2006; Schlosser 2009; and for critical discussion Sober 1999b; Buckareff 2011, 2012).

[29] See Burge (2007) and Woodward (2008a, 244–249) for arguments against some of these lines of objection.

explain, they must be either weakly or strongly sufficient for their effects. It is unclear to me why the exclusion problem is often framed so that mental causes must be sufficient for their effects in a stronger sense than sufficiency in the circumstances. Bennett (2003, Section 5) thinks that anything less than sufficiency would endanger the 'full-fledged causal efficacy of the mental' (481), granting it merely 'a derivative efficacy' (482). I cannot see the motivation for claims of this form if sufficiency is supposed to be stronger than circumstantial sufficiency—especially given the metaphors that are often used to characterise the exclusion problem.[30] If an event has a complete physical cause, mental causes are often said to have 'no work left to do' (Kim 1998, 35, 37, 54, 110, 126 n. 6), 'no gaps left to fill' (Menzies 2003), no opportunity to 'inject themselves' into the causal order (Kim 1998, 41; 2005, 16); if there is no lowest level of causation, we are supposed to worry that causal powers will 'drain away' (Block 2003; Kim 2003). But if there *was* work left to do, a gap to be filled, an injection to be provided, or a drain to be plugged, presumably the context would be almost sufficient, and the additional impetus plus context would be wholly sufficient. The work, filler, injection, or plug would not itself be wholly sufficient, but rather would be sufficient in the circumstances. Now perhaps these are all just poor metaphors for what is supposed to be at issue here; but metaphors aside, the claim in question would be that any actual causes that are not at least weakly sufficient must have merely derivative efficacy. Given that sufficiency in the circumstances is the sort of efficacy most causes have in most scientific theories, I say that derivative causes in this idiosyncratic sense would be causes enough for mental causation.[31]

3.4 Exclusion

It may seem that the formulation of **Exclusion** is now straightforward: it should simply be the strongest principle that is inconsistent with **Non-Reductionism**$_j$, **Completeness**$_j$, and **Mental Causation**$_j$. However, for the principle to have any *prima facie* plausibility, it needs to be weaker. To translate a point first made by Goldman (1969, 470–473; 1970, 159–161) into this context, **Completeness**$_j$ is perfectly compatible with **Mental Causation**$_j$ in a case where there is a path from the mental variable that is an actual cause of the event to the physical variable that is strongly sufficient for the event, or in a case where the mental variable is on a path from the physical variable to the event. A principle that says that if an event has a sufficient cause it has no other causes is clearly far too strong to be plausible, for it is inconsistent with the existence of causal chains.

[30] I do not suggest Bennett endorses the position I here criticise.

[31] For a more detailed argument for this claim, see Woodward (2008a, 245–249).

My final formulation of **Exclusion** is therefore:

- **Exclusion**$_j$. If there exists a causal model with variables drawn from a vocabulary F, which is strongly closed with respect to variables drawn from vocabulary G, and in which a variable F_1 is a strongly sufficient actual cause for an event, then there exists no causal model in which a variable G_1 drawn from vocabulary G is an actual cause of that event, unless there is a causal chain containing F_1 and G_1.

An attractive feature of this formulation is that it reveals a way in which someone who rejects **Non-Reductionism**$_j$ can evade the exclusion problem. Here I have in mind Lowe (2000, 2003), who has argued that all causal closure principles with strong empirical support are logically consistent with non-physicalist theories on which mental causes occupy a place in causal chains *between* sufficient physical causes and their effects. While I think we have overwhelmingly strong reasons to reject theories of this sort, my formulations of **Completeness**$_j$, **Exclusion**$_j$, and **Mental Causation**$_j$ support Lowe's claim.

The non-reductive physicalist, on the other hand, is in no position to make a similar move. They would thereby be committed to a position on which mental causes occupy a place in causal chains between sufficient physical causes and their effects, but are metaphysically necessitated by *different* physical variables, which *are not themselves sufficient for those effects*. It is rare to find a position in logical space no philosopher is willing to occupy, but this must be one of them.[32] My final formulation of **Mental Causation**$_j$ is therefore the following, which closes this loophole and renders the propositions that form the exclusion problem logically inconsistent:

- **Mental Causation**$_j$. There exists a causal model with variables drawn from M in which a mental variable M_1 is an actual cause of an event, and there is no causal chain containing P_1 and M_1.

3.5 The Interventionist Exclusion Problem

Putting this all together, I conclude that the exclusion problem in the interventionist framework should be formulated as follows:

- **Non-Reductionism**$_j$. Mental variables are distinct from physical variables in the sense that they are drawn from distinct vocabularies M and P, and the values of the M-variables are metaphysically necessitated by the values of the P-variables.

[32] See Kim (1998, 37, 40, 44). As Kim notes, the non-reductive physicalist will invariably be committed to versions of physicalism and closure on which this option is ruled out.

- **Completeness$_j$**. For every event, there exists a causal model with variables drawn from P, which is strongly closed with respect to M, in which there is a strongly sufficient actual cause P_1 for that event.
- **Mental Causation$_j$**. There exists a causal model with variables drawn from M in which a mental variable M_1 is an actual cause of an event, and there is no causal chain containing P_1 and M_1.
- **Exclusion$_j$**. If there exists a causal model with variables drawn from a vocabulary F, which is strongly closed with respect to variables drawn from vocabulary G, and in which a variable F_1 is a strongly sufficient actual cause for an event, then there exists no causal model in which a variable G_1 drawn from vocabulary G is an actual cause of that event, unless there is a causal chain containing F_1 and G_1.

One of these claims must go. We are finally in a position to consider arguments for **Exclusion$_j$**.

4 Compatibilism Examined

In this section I evaluate two arguments for **Exclusion$_j$**. The first is familiar from discussion of the exclusion problem in the interventionist setting, but the second is not. This is because the discussion has almost universally assumed a particular conception of the relationship between physical and mental variables, according to which the mental cause M_1 that figures in **Mental Causation$_j$** is metaphysically necessitated by the strongly sufficient actual cause P_1 that figures in **Completeness$_j$**. I will call this assumption *subvenience sufficiency*, the non-reductive physicalist position that accepts it *internalism*, and the non-reductive physicalist position that denies it *externalism*. As I will show, the distinction is important. I begin with a discussion of subvenience sufficiency itself, and then consider internalism and externalism in turn. I side with those who take the interventionist to have a good response to the argument for **Exclusion$_j$** under the assumption of internalism. But I go on to argue that the response the interventionist has to the argument for **Exclusion$_j$** under the assumption of externalism serves to expose a weakness in interventionism itself.

4.1 Subvenience Sufficiency

As formulated, the exclusion problem invites us to consider two causal models. Each model contains a variable E_1 that is a candidate effect for a mental cause. The first model, $\mathcal{M}_P$, the existence of which is entailed by **Completeness$_j$**, contains (in addition to E_1) only variables drawn from P, is strongly closed with respect to M, and contains a strongly sufficient actual cause P_1 for E_1. The second model, $\mathcal{M}_M$, the existence of which is entailed by **Mental Causation$_j$**, contains (in addition to

E_1) variables drawn from M, and contains an actual cause M_1 of E_1. As is also entailed by **Mental Causation**$_j$, there is no causal chain containing P_1 and M_1. I will make the simplifying assumptions that $\mathcal{M}_M$ contains (in addition to E_1) only variables drawn from M, and that both $\mathcal{M}_P$ and $\mathcal{M}_M$ only contain variables on paths terminating in variable E_1.

With respect to models of this sort, subvenience sufficiency can be defined as follows:

- **Subvenience Sufficiency.** Given two models $\mathcal{M}_F$ and $\mathcal{M}_G$, where each model only contains variables on paths terminating in variable E_1, $\mathcal{M}_F$ is *subvenience sufficient* with respect to $\mathcal{M}_G$ and E_1 *iff* the values of all other variables in $\mathcal{M}_G$ are metaphysically necessitated by the values of strongly sufficient causes of E_1 in $\mathcal{M}_F$.

It is important to see that $\mathcal{M}_P$ being subvenience sufficient with respect to $\mathcal{M}_M$ is a substantive assumption that is not itself entailed by **Non-Reductionism**$_j$ either alone or in conjunction with **Completeness**$_j$ and **Mental Causation**$_j$. In particular, while **Non-Reductionism**$_j$ merely requires that the values of the M-variables are metaphysically necessitated by the values of the P-variables, $\mathcal{M}_P$ being subvenience sufficient with respect to $\mathcal{M}_M$ imposes the much stronger constraint that the M-variables are metaphysically necessitated by the very P-variables that are strongly sufficient for their effects. The assumption is vividly illustrated by what Loewer (2015, 60) calls 'Kim's Favourite Diagram' (2003, 159), a way to represent the exclusion argument that is ubiquitous in Kim's work, in which one and the same physical event is represented as both the cause of a given effect, and as the subvenience basis for the mental event which Kim takes it to exclude (see Figure 4.2).

I will refer to non-reductive physicalism in conjunction with **Subvenience Sufficiency** as *internalism* and non-reductive physicalism without **Subvenience Sufficiency** as *externalism*. The fact that internalism has been so frequently assumed in the discussion of the exclusion problem would be unremarkable if it were not the case that most non-reductive physicalists are committed to rejecting it, and if it did not make a difference to the arguments available to the non-reductive

$$\begin{tikzcd} \mathrm{M} & \mathrm{M}^* \\ \mathrm{P} \arrow[u, "\text{Supervenes}"] \arrow[r, "\text{causes}"] & \mathrm{P}^* \arrow[u, "\text{Supervenes}"'] \end{tikzcd}$$

Figure 4.2 Kim's Favourite Diagram.

physicalist for rejecting **Exclusion**$_j$. As Bennett (2003) notes, internalism is inconsistent with content externalism, with functionalism in general, and with conceptual role semantics in particular.[33] These are the most prominent of the theories of mental properties that motivate non-reductionism in the first place, so it is hardly open to the non-reductive physicalist to ignore their consequences. I begin, however, with internalism.

4.2 Internalism

Baumgartner (2009, 2010) has argued that if interventionism is true, then whenever variables stand in relationships of metaphysical necessitation, the necessitated variable cannot have any of the same effects as the necessitating variable. If that is right, then internalism is incoherent. For the internalist is committed, by virtue of the claim that $\mathcal{M}_P$ is subvenience sufficient with respect to $\mathcal{M}_M$, to the existence of variables that stand in exactly this relationship. In this section I argue that the correct formulation of interventionism blocks this argument.[34]

As was clear from the definitions introduced in Section 2, all of the fundamental interventionist causal concepts are defined in terms of interventions. Baumgartner's argument depends on the way in which interventions are defined by Woodward (2003, 98). Woodward first introduces the type-level notion of an *intervention variable*:

- **IV.** I is an *intervention variable* for X with respect to Y *iff*:
  - **I**$_1$. I is a contributing cause of X.
  - **I**$_2$. There is a model in which I has at least one value that is weakly sufficient for the value of X.
  - **I**$_3$. Every causal chain from I to Y contains X.
  - **I**$_4$. I is statistically independent of every contributing cause of Y on causal chains that do not contain X.

This is then used to define the token-level notion of an *intervention*:

- **IN.** $I = i$ is an *intervention* on X with respect to Y *iff* is an intervention variable for X with respect to and there is a model in which $I = i$ is a weakly sufficient cause of the value of X.

[33] In addition, Worley (1993, Section 5) argues that internalism is inconsistent with anomalous monism and the folk-psychological platitude that a single mental state may be a cause of different effects on different occasions.

[34] I am indebted here to correspondence with Michael Baumgartner.

Note that I have presented definitions that are weaker than Woodward's in one respect, since $\mathbf{I}_2$ is weaker than Woodward's condition. The distinction amounts to whether interventions must be *hard*, so that they override all other causal connections to the variable intervened on, or whether they may be *soft*, and merely make an additional causal impact to the variable intervened on. I opt for the weaker definitions because Baumgartner's argument works either way, and because Woodward himself accepts both formulations (2015b, 3584; 2015a, 321 fn. 15; 2017, 254 fn. 3).[35]

Baumgartner's argument is simple, with each premise following from the definitions of the relevant notions, or from claims the internalist is committed to accepting. For M_1 to be an actual cause of E_1 in $\mathcal{M}_M$, it must be a contributing cause of E_1 in $\mathcal{M}_M$ ($\mathbf{AC}_A$). For it to be a contributing cause of E_1 in $\mathcal{M}_M$, there must be an intervention on M_1 with respect to E_1 (**CC**). For there to be an intervention on M_1 with respect to E_1, there must be an intervention variable I for M_1 with respect to E_1 (**IN**). For there to be an intervention variable I for M_1 with respect to E_1, every causal chain from I to E_1 must contain M_1 (**IV**, $\mathbf{I}_3$), and I must be statistically independent of every contributing cause of E_1on causal chains that do not contain M_1 (**IV**, $\mathbf{I}_4$). But internalism is committed to the claim that $\mathcal{M}_P$ is subvenience sufficient with respect to $\mathcal{M}_M$. This entails that there is no way to make a change to M_1 without also changing P_1, which in turn entails both that there is a causal chain from I to P_1 to E_1 that does not contain M_1, and that P_1 is not statistically independent of M_1. So there is no such intervention variable I, M_1 is not an actual or contributing cause of E_1, and there is no such model $\mathcal{M}_M$ as required by the internalist.[36]

In response to this argument, Woodward (2015a, 323) helpfully distinguishes three questions. First, are the definitions that lead to this result adequate interpretations of Woodward (2003)? Second, must any interventionist theory that deserves the name adopt definitions that lead to this result? Third, in order for variables to be causes, must they make a difference to their effects beyond the differences made by variables that metaphysically necessitate them? I set Woodward's first question aside.[37] On Woodward's second question, some authors have considered ways in which the basic interventionist framework can be expanded, so that variables related both causally and by metaphysical necessitation can appear in the same model. The debate then becomes whether the principles that should

[35] For discussion of hard and soft interventions, see Korb et al. (2004); Markowetz, Grossmann, and Spang (2005); Eberhardt and Scheines (2007); Eberhardt (2014), and in the psychological context Campbell (2007); Kaiserman (2020).

[36] As Baumgartner (2009, 171) notes, the argument does not depend on a premise concerning causal closure: it can be used to show that, on these definitions, no variables related by metaphysical necessity can share any effects. Gebharter (2017a) argues that this is also the case with respect to the argument in Gebharter (2017b), and Stern and Eva (2023) agree.

[37] Woodward (2015a, 324–325; 2017, 257) has argued that the answer is 'no'. As he says, it is the least interesting of the three.

govern models of this sort should permit or prohibit causation by necessitated variables (Baumgartner 2010; Woodward 2015a; Gebharter 2017b; Stern and Eva 2023). If we wish to restrict our focus to causal relationships, however, there exists a more conservative amendment of the interventionist framework that is sufficient to block Baumgartner's argument.[38]

The amendment is as follows:

- **IV***. *I* is an *intervention variable* for *X* with respect to *Y iff*:
  - **I_1**. *I* is a contributing cause of *X*.
  - **I_2**. There is a model in which *I* has at least one value that is weakly sufficient for the value of *X*.
  - **I_3^***. Every path from *I* to *Y* goes through *X* in every model containing *I*, *X* and *Y*.
  - **I_4^***. *I* is statistically independent of every contributing cause of *Y* on paths that do not contain *X* in every model containing *I*, *X* and *Y* .

The difference between **IV** and **IV*** concerns the third and fourth conditions. Condition **I_3** requires that there are no paths, in any model, from *I* to *Y* without *X* . Condition **I_3^*** relaxes this requirement, requiring that there are no paths from *I* to *Y* without *X in any model that contains those variables*. Likewise, condition **I_4** requires that *I* is statistically dependent of contributing causes of *Y*, in all models, that are on paths without *X* . **I_4^*** relaxes this requirement, requiring that *I* is statistically dependent of contributing causes of *Y*, on paths without *X, in any model that contains those variables.*

The difference between these definitions shows up as a consequence of **Independent Manipulability** (Section 3.1), according to which it is a condition on variables coexisting in a model that it be metaphysically possible for every proper subset to be set to every combination of their possible values by independent interventions. This entails:

- **Non-Necessitation**. A causal model cannot contain a variable with a possible value that is metaphysically necessitated by a possible combination of values of any proper subset of the other variables in the model.

An immediate consequence of this, if internalism is true, is that there are no causal models that contain both P_1 and M_1. This in turn blocks Baumgartner's argument. According to **IV***, the presence of a causal chain from *I* to P_1 to E_1 that does not contain M_1, and the fact that P_1 is not statistically independent of M_1, do not threaten the satisfaction of **I_3^*** and **I_4^***. More generally, the fact that any change to M_1

[38] The basic strategy I develop in the remainder of this section is also proposed by Eronen and Brooks (2014), who cite an earlier version of this paper. I do not endorse their arguments for it.

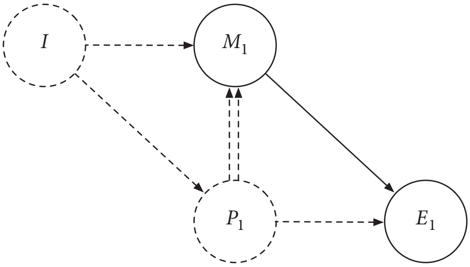

Figure 4.3 Diagram for $\mathcal{M}_M$, if internalism is true. Solid lines represent variables and direct causes that are in the model. Dashed lines represent variables and direct causes that are not in the model. Double arrows represent metaphysical necessitation. According to **IV**, I cannot be an intervention variable for M_1 with respect to E_1. According to **IV***, it can.

entails some corresponding change to P_1 is no obstacle to there being well-defined interventions on M_1 (see Figure 4.3).

The answer to Woodward's second question, therefore, is 'no'. There is a coherent formulation of an interventionist theory of causation that does not generate the consequence that a necessitated variable cannot have any of the same effects as the necessitating variable. Moreover, it is a formulation that is perfectly suited to the non-reductive physicalist, in the following sense. As Bennett (2008) argues, the non-reductive physicalist would ideally like a solution to the exclusion problem that can play two roles. On the one hand, it should show that causal considerations do not force her into reductive physicalism. On other hand, it should show that causal considerations are still a problem for the dualist.[39] Interventionism formulated in terms of **IV*** has both consequences, at least for the internalist. On the one hand, as I have just argued, the internalist can argue that their position is coherent, and compatible with mental causation. On the other hand, the internalist can point out that the dualist, in virtue of rejecting the metaphysical necessitation of the values of mental variables by the values of physical variables, cannot avail themselves of the same sorts of interventions on mental variables. Instead, they must commit to the existence of interventions on mental variables that do not entail any changes to physical variables, and that are statistically independent of physical variables. But in that case, accepting **Completeness**$_j$ would force the dualist to admit that while there may be such interventions, they could not result in any downstream effects. As a result, the dualist who accepts **Completeness**$_j$ is

[39] In the current framework, *reductive physicalism* can be defined as the position on which mental variables are not distinct from physical variables, and *dualism* can be defined as the position on which mental variables are distinct from physical variables, and the values of the mental variables are not metaphysically necessitated by the values of the physical variables.

committed to rejecting **Mental Causation**$_j$. Here is a different way to see the point. **Completeness**$_j$ says that $\mathcal{M}_p$ is strongly closed with respect to M. This means that there is no model containing all of the variables from $\mathcal{M}_p$, and any variables from M, in which any strongly sufficient causes lose that status. But there are two ways this can be true. The first is for there to be no such model containing all of the variables from $\mathcal{M}_p$ and any variables from M. This is what internalism is committed to, and it is consistent with the existence of a model in which M_1 is a cause. The second is for there to be such a model. This is what dualism is committed to, and it is not consistent with M_1 being a cause, in that model or any other.[40]

I turn now to Woodward's third question. What can be said to recommend interventionism formulated in terms of **IV*** over interventionism formulated in terms of **IV**, besides the fact that it facilitates a coherent non-reductionism?

One argument derives from the very motivation for the theory. If there is a single idea at the heart of interventionism, it is that the best way to understand the nature of causation is to theorise through the lens of an ideal experiment for detecting it (Woodward 2003, 14). So for example, when I introduced **Independent Manipulability** (Section 3.1), I said that it is motivated by the idea that for variables to coexist in a model, it must be possible to non-trivially test, for every pair, whether they are related by direct causation. The same motivation can be given for **IV***. The idea behind conditions $\mathbf{I}_3$ and $\mathbf{I}_4$ is that interventions should be independent of potential confounding causes. But just as the interventionist should say that variables are only candidates for being causally related if they can be independently manipulated, they should say that variables are only candidates for being potential confounding causes if they can be independently manipulated. $\mathbf{I}_3$ and $\mathbf{I}_4$ do not entail this constraint, but $\mathbf{I}_3^*$ and $\mathbf{I}_4^*$ do.

This is not a new form of argument. In his discussion of causal completeness in the context of probabilistic theories of causation, Sober (1999a, Section 2) considers theories according to which a positive causal factor must raise the probability of an effect in at least one background context, and lower it in none:

- **Positive Causal Factor**. C is a *positive causal factor* for E *iff* $P(E \mid C \,\&\, X_i) \geq P(E \mid \neg C \,\&\, X_i)$ for all background contexts X_i, with strict inequality for at least one X_i.

What counts as a background context? According to Sober, a necessary condition on a set of properties constituting a background context relative to a given cause and

[40] Here I disagree with Shapiro and Sober (2007), who suggest that a well-conceived argument for epiphenomenalism, under the assumption of interventionism, 'should aim to show that one class of properties does not affect a second class, not that the first has no effects at all' (241). This underestimates how strong the constraints are that completeness principles put on the sorts of properties that can be causes.

effect is that these probabilities are well defined. As he notes, this entails that when evaluating whether a necessitated property is a causal factor, necessitating properties cannot be part of any background context, since then $P(\neg C \,\&\, X_i)$ would be 0 and $P\left(E \mid \neg C \,\&\, X_i\right)$ not well defined. Sober's argument is identical to the argument I have just given for **IV***, transposed to the probabilistic case.[41]

The convergence of these arguments underscores that the issue concerning exclusion for difference-making theories of causation, under the assumption of internalism, concerns the contexts relative to which causes must make a difference. Must they make a difference controlling for *all* other causes, or must they make a difference controlling for all other *independent* causes? (Shapiro and Sober 2007, 241). I will briefly describe three other arguments that can be given for the second conception, before turning to externalism.

First, it is implicit in scientific practice that you do not need to control for necessitating variables in order to be justified in believing that necessitated variables are causes (Shapiro and Sober 2007; Shapiro 2010). As Sober puts it: 'This fact about scientific practice stands on its own' (1999a, 147). In this connection, it is also important to note that allowing necessitated properties to be causes does not mean that they trivially satisfy the requirements to be causes, simply in virtue of their being necessitated by causes (Segal and Sober 1991; Shapiro and Sober 2007, 256–259; Woodward 2015a, 2017). It is a substantive and difficult matter to determine whether a necessitated property meets the conditions for causation by the lights of interventionism under the assumption of **IV***.

Second, it is clear that in examples involving logical or conceptual necessitation between variables we do not need to hold fixed one variable in order to determine whether the other makes a difference (Woodward 2008a, 2015a). Indeed, since for any variable we can introduce others related to it in these ways, imposing this requirement would mean that no variables could possibly satisfy the requirements for being causally related. It can then be argued either that metaphysical necessitation is relevantly similar to those forms of dependence, or that **IV*** provides the correct theory in light of that fact.

Third, it can be argued that **IV** is, but **IV*** is not, subject to the argument that if there were no fundamental causal level, causation would drain away (Block 2003; Kim 2003).

I do not claim that these arguments are collectively decisive. But I do claim that in interventionism formulated with **IV***, the non-reductionist has a coherent and well-motivated theory of causation that entails the falsity of **Exclusion**$_j$. If the

[41] The same line of thought is arguably implicit in Eells (1991, 31). Similarly, Humphreys (1989, 74) requires that it be physically possible for the cause and its absence to occur relative to all background factors (for an application to the exclusion problem, see Henderson 1994). An analogous argument, in the context of a theory of causation along the lines of Mackie (1974), is given by Melnyk (2003, 137–138).

non-reductionist is an internalist, there is no obstacle to their endorsing **Mental Causation**$_j$.

4.3 Externalism

As it happens, very few non-reductionists can rest content at this point. For most of the conceptions of mental properties that motivate non-reductionism in the first place entail that internalism is false. So we need to consider arguments that target the externalist conception of mental properties. I will begin by discussing the externalist position generally, and then discuss some of the more concrete forms it may take when they become relevant.

The first point to note is that the externalist cannot make use of the same line of reasoning available to the internalist, who can appeal to **Non-Necessitation** in order to argue that there is no model containing both M_1 and P_1. Since the externalist by definition rejects the necessitation of M_1 by P_1, there is no obstacle to the existence of a model that contains both variables. Since the externalist remains a physicalist, they must thereby be committed to the existence of other physical variables that, together with P_1, necessitate M_1. For simplicity, I will use a single variable P_2 to represent these. So the externalist is committed to P_1 and P_2 together necessitating M_1, and neither P_1 nor P_2 alone necessitating M_1.

It now appears that epiphenomenalism looms. Consider model $\mathcal{M}_{PM1}$, containing variables P_1, M_1, and E_1. As I noted in Section 3.2, **Completeness**$_j$ entails that there is no path from M_1 to E_1 in $\mathcal{M}_{PM1}$. Moreover, this is not because there cannot be an intervention variable for M_1 with respect to E_1. There can be, but it must involve changing M_1 by changing P_2 (which is permissible according to **IV***). However, a difference of that sort cannot make any additional difference to E_1. At least in model $\mathcal{M}_{PM1}$, M_1 is epiphenomenal (see Figure 4.4).[42]

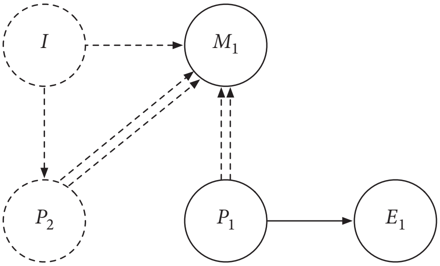

Figure 4.4 Diagram for $\mathcal{M}_{PM1}$.

[42] Arguments of this form are discussed by Block (1990), Worley (1993), and Rescorla (2012, 2014, Section 7).

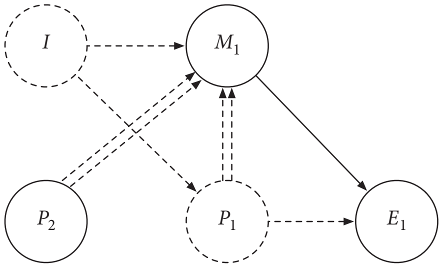

Figure 4.5 Diagram for $\mathcal{M}_{PM2}$.

Can we conclude that M_1 is epiphenomenal *simpliciter*? Not without additional argument. For according to the definitions provided in Section 2, any immediate inference from a variable not causing another in a model to its not causing it *simpliciter* is invalid. Recall that $X = x$ is an actual cause of $Y = y$ *simpliciter iff* there exists a model in which $X = x$ is an actual cause of $Y = y$. It follows that there is an asymmetry in what it takes to show that a variable value is or is not an actual cause of another. To show that a variable value *is* an actual cause, we simply need to identify a model in which it is. But to show that a variable value *is not* an actual cause, we need to show that there *does not exist* a model in which it is. What is needed to establish **Exclusion**$_j$ under the assumption of externalism is an argument that could establish the non-existence of a model in which M_1 is an actual cause of E_1, on the basis that it is not an actual cause in $\mathcal{M}_{PM1}$.

In fact, the externalist can do better than simply rejecting this inference. For they can exhibit a model in which M_1 is a cause of E_1. Consider model $\mathcal{M}_{PM2}$, containing variables P_2, M_1, and E_1. In this model, if we hold P_2 fixed at some particular value, then any intervention on M_1 must change P_1. So long as there exists a change of this sort that is associated with a change to E_1, then M_1 will be a direct cause of E_1, and for at least one state of the model will be an actual cause of E_1 (see Figure 4.5). Interventionism is therefore consistent not only with mental causation under the assumption of internalism, but with mental causation under the assumption of externalism.

For illustration, consider the case of content externalism. Here P_1 can be interpreted as representing neurophysiological properties, P_2 can be interpreted as representing content-fixing environmental properties, and M_1 can be interpreted as representing externally individuated mental properties, where the values of M_1 are metaphysically necessitated by the values of P_1 and P_2. $\mathcal{M}_{PM1}$ reveals the fact that if we hold fixed the neurophysiological properties, altering mental properties by altering the content-fixing environmental properties on which they partly depend would make no difference to behaviour. $\mathcal{M}_{PM2}$, on the other hand, reveals the fact that if we hold fixed the content-fixing environmental properties, altering mental properties by altering the neurophysiological properties on which they partly depend may make a difference to behaviour. According to interventionism, the existence of the former model does not entail that mental properties are not causes of behaviour, while the existence of the latter model entails that mental properties are causes of behaviour. If the non-reductionist is an externalist, there is no obstacle to their endorsing **Mental Causation**$_j$.

4.4 A Weakness in Interventionism

I have argued that interventionism allows both internalists and externalists to consistently accept **Non-Reductionism**$_j$, **Completeness**$_j$, and **Mental Causation**$_j$. In other words, an interventionist is entitled to reject **Exclusion**$_j$. However, in this section I suggest that attention to the externalist case reveals a weakness in interventionism.

The basic form of the problem is identified by Rescorla (2014, Section 11). As Rescorla notes, there are situations in which structurally identical models to $\mathcal{M}_{PM1}$ and $\mathcal{M}_{PM2}$ apply, and yet in which it is not the case that M_1 is a cause of E_1. Take, for example, a simple pocket calculator (Haugeland 1985, 121–123; Rescorla 2014, 180–181). The semantic properties instantiated by the calculator during the course of a calculation (let these be represented by M_1) are jointly determined by two factors: the physical properties it instantiates (let these be represented by P_1) and the interpretation to which they are subject (let this be represented by P_2). In this context, claims parallel to those concluding the previous section can now be introduced. Consider some particular output of the calculator (let this be represented by E_1). $\mathcal{M}_{PM1}$ reveals the fact that if we hold fixed the physical properties of the calculator, altering its semantic properties by altering the interpretation to which its physical properties are subject would make no difference to the output. $\mathcal{M}_{PM2}$, on the other hand, reveals the fact that if we hold fixed the interpretation to which its physical properties are subject, altering its semantic properties by altering its physical properties may make a difference to the output. But semantic properties don't cause the outputs of pocket calculators (Rescorla 2012). Something has gone wrong.

Moreover, the problem cannot be evaded by simply rejecting externalism. For there are many other situations in which structurally identical models to $\mathcal{M}_{PM1}$ and $\mathcal{M}_{PM2}$ apply, and in which M_1 is a cause of E_1. For example, consider a match struck in the presence of air, causing it to light. Let the presence of air be represented by M_1, the presence of oxygen be represented by P_1, the presence of all other constituents of air be represented by P_2, and the match lighting be represented by E_1. $\mathcal{M}_{PM1}$ reveals the fact that if we hold fixed the presence of oxygen, altering the presence of air by altering the presence of the other constituents would make no difference to the match lighting. $\mathcal{M}_{PM2}$, on the other hand, reveals the fact that if we hold fixed the presence of the other constituents, altering the presence of air by altering the presence of oxygen would make a difference to the match lighting.[43]

In sum, M_1 is a cause of E_1 in only some of the cases in which it appears that $\mathcal{M}_{PM2}$ applies, and the interventionist therefore owes us an account both of the difference between the cases, and why we should believe that mental properties fall on the right side of the line.[44]

5 Conclusion

It has been more than 25 years since the publication of Jaegwon Kim's *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation* (1998), the canonical investigation of causal exclusion principles. In summarising his discussion of counterfactual theories of causation, Kim wrote (71–72):

> […] what the counterfactual theorists need to do is to give an *account* of just what makes those mind-body counterfactuals we want for mental causation true, and show that on that account those counterfactuals we don't want, for example, epiphenomenalist counterfactuals, turn out to be false. Merely to point to the apparent truth, and acceptability, of certain mind-body counterfactuals as a vindication of mind-body causation is to misconstrue the philosophical task at hand … Such gestures only show that mind-body causation is part of what we normally take to be the real world; they go no further than a mere reaffirmation of our belief in the reality of mental causation. What we want—at least, what some of us are looking for—is a philosophical account of *how* it can be real in light of other principles and truths that seem to be forced upon us.

The 'principles and truths' Kim refers to here are those that he took as the premises in his arguments for causal exclusion principles. I have argued that an interventionist is entitled to reject those principles. But I have also argued that interventionists have not yet discharged the obligations that Kim here describes. In particular, they need to explain what is defective about the application of $\mathcal{M}_{PM2}$

[43] For discussion of this example in the context of mental causation, see Segal and Sober (1991), Tye (1991), Peacocke (1993a, 1993b), and Segal (2004, 2009). Note that examples of this sort place pressure on a condition Woodward (2008a, 2021a, 2021b, 2022) calls *realisation independence*, which requires that interventions must have the same effect no matter how they are realised. This condition seems to entail that in any case in which structurally identical models to $\mathcal{M}_{PM1}$ and $\mathcal{M}_{PM2}$ apply, M_1 is not a cause of E_1. See Hoffman-Kolss (2014, Section 5) for a different argument against realisation independence.

[44] A referee for this volume suggests that the notion of *conditional independence* recently discussed by Woodward (2021c, 2020, 2021a, 2021b, 2022) may help here. I do not think so, for two reasons. First, and in my view correctly, Woodward does not propose that conditional independence is a necessary condition on causation. Second, nothing I have said entails whether or not conditional independence is satisfied in either the case of mental properties or the case of the calculator, and I do not see any principled reason for saying it must always hold in the former and never in the latter.

to the case of the pocket calculator. Work of this sort must proceed along two paths: the development of principled constraints on when a causal model is appropriate for a given situation, as in, for example, Hitchcock (2001, 2004, 2012), Halpern and Hitchcock (2010), Halpern (2016), Woodward (2016), Blanchard and Schaffer (2017), McDonald (forthcoming, 2023), and Hoffmann-Kolss (this volume); and the application of these constraints to specific conceptions of the relationship between physical and mental properties, as in, for example, Rescorla (2014). Only when this cumulative case has been made, for the difference between pocket calculators and minds, can an interventionist claim to have a fully principled basis for rejecting **Exclusion.**[45]

References

Baumgartner, Michael. 2009. "Interventionist Causal Exclusion and Non-Reductive Physicalism." *International Studies in the Philosophy of Science* 23 (2): 161–178. http://doi.org/10.1080/02698590903006909.

Baumgartner, Michael. 2010. "Interventionism and Epiphenomenalism." *Canadian Journal of Philosophy* 40 (3): 359–383. https://doi.org/10.1080/00455091.2010.10716727.

Baumgartner, Michael. 2013. "Rendering Interventionism and Non-Reductive Physicalism Compatible." *Dialectica* 67 (1): 1–27. http://doi.org/10.1111/1746-8361.12008.

Baumgartner, Michael. 2018. "The Inherent Empirical Underdetermination of Mental Causation." *Australasian Journal of Philosophy* 96 (2): 335–350. https://doi.org/10.1080/00048402.2017.1328451.

Bennett, Karen. 2003. "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It." *Noûs* 37 (3): 471–497. http://doi.org/10.1111/1468-0068.00447.

Bennett, Karen. 2008. "Exclusion Again." In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 280–305. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.003.0015.

Blackburn, Simon. 1991. "Losing Your Mind: Physics, Identity, and Folk Burglar Prevention." In *The Future of Folk Psychology: Intentionality and Cognitive Science*, edited by John D. Greenwood, 196–225. Cambridge: Cambridge University Press.

Blanchard, Thomas, and Jonathan Schaffer. 2017. "Cause Without Default." In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 175–214. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198746911.001.0001.

Block, Ned. 1990. "Can the Mind Change the World?" In *Meaning and Method: Essays in Honor of Hilary Putnam*, edited by George Boolos, 137–170. Cambridge: Cambridge University Press.

[45] Earlier versions of this paper were presented to a Mellon Mental Causation Workshop at Syracuse University in 2007, to the CUNY Cognitive Science Symposium in 2008, and to a workshop at Harvard University in 2019. I am grateful to the audiences on those occasions, and especially to Karen Bennett, Tony Dardis, David Rosenthal, Samuel Lee, and Ned Hall. I also thank Arif Ahmed, Ben Blumson, David Braddon-Mitchell, Michael Brent, John Campbell, Brandon Carey, Shamik Dasgupta, Sommer Hodson, Adam Kay, Doug Kutach, David Macarthur, Kevin McCain, Peter Menzies, Huw Price, and Nandi Theunissen for discussion or feedback, and Michael Baumgartner and Michael Rescorla for very helpful correspondence, on earlier versions. More recently, I thank Michael Tye for sending me his work on this topic, and Alastair Wilson, Katie Robertson, Henry Taylor, Mike Hicks, Ben Curry, and two referees for this volume for their advice on preparing the version that appears in this volume.

Block, Ned. 2003. "Do Causal Powers Drain Away?" *Philosophy and Phenomenological Research* 67 (1): 133–150. http://doi.org/10.1111/j.1933-1592.2003.tb00029.x.

Bontly, Thomas D. 2002. "The Supervenience Argument Generalizes." *Philosophical Studies* 109 (1): 75–96. http://doi.org/10.1023/A:1015786809364.

Buckareff, Andrei A. 2011. "Intralevel Mental Causation." *Frontiers of Philosophy in China* 6 (3): 402–25. http://doi.org/10.2307/44259314.

Buckareff, Andrei A. 2012. "An Action-Theoretic Problem for Intralevel Mental Causation." *Philosophical Issues* 22: 89–105. http://doi.org/10.2307/41683062.

Burge, Tyler. 1993. "Mind-Body Causation and Explanatory Practice." In *Mental Causation*, edited by John Heil and Alfred Mele, 97–120. Oxford: Oxford University Press.

Burge, Tyler. 2007. "Postscript: Mind-Body Causation and Explanatory Practice." In *Foundations of Mind*, 2: 363–382. Philosophical Essays. Oxford: Oxford University Press.

Campbell, John. 2007. "An Interventionist Approach to Causation in Psychology." In *Causal Learning: Psychology, Philosophy, Computation*, edited by Alison Gopnik and Laura Schulz, 58–66. New York: Oxford University Press.

Campbell, John. 2007. 2008. "Causation in Psychiatry." In *Philosophical Issues in Psychiatry: Explanation, Phenomenology and Nosology*, edited by Kenneth S. Kendler and Josef Parnas, 196–215. Baltimore MD: Johns Hopkins University Press.

Cartwright, Nancy. 1983. *How the Laws of Physics Lie*. Oxford: Oxford University Press. http://doi.org/10.1093/0198247044.001.0001.

Cartwright, Nancy. 1994. "Fundamentalism Vs. The Patchwork of Laws." *Proceedings of the Aristotelian Society* 94 (1): 279–292. https://doi.org/10.1093/aristotelian/94.1.279.

Crane, Tim. 2008. "Causation and Determinable Properties: On the Efficacy of Colour, Shape, and Size." In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 176–195. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.003.0011.

Davidson, Donald. 1963. "Actions, Reasons, and Causes." *Journal of Philosophy* 60 (23): 685–700. http://doi.org/10.2307/2023177.

Davidson, Donald. 1967. "Causal Relations." *Journal of Philosophy* 64 (21): 691–703. https://doi.org/10.2307/2023853.

Davidson, Donald. 1970. "Mental Events." In *Experience and Theory*, edited by Lawrence Foster and Joe William Swanson, 79–101. Amherst MA: University of Massachusetts Press. http://doi.org/10.1093/0199246270.003.0011.

Davidson, Donald. 1995. "Laws and Cause." *Dialectica* 49 (2–4): 263–279. https://doi.org/10.1111/j.1746-8361.1995.tb00165.x.

Eberhardt, Frederick. 2014. "Direct Causes and the Trouble with Soft Interventions." *Erkenntnis* 79 (4): 755–777. https://doi.org/10.1007/s10670-013-9552-2.

Eberhardt, Frederick, and Richard Scheines. 2007. "Interventions and Causal Inference." *Philosophy of Science* 74 (5): 981–995. https://doi.org/10.1086/525638.

Eells, Ellery. 1991. *Probabilistic Causality*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511570667.

Eronen, Markus I., and Daniel S. Brooks. 2014. "Interventionism and Supervenience: A New Problem and Provisional Solution." *International Studies in the Philosophy of Science* 28 (2): 185–202. https://doi.org/10.1080/02698595.2014.932529.

Fenton-Glynn, Luke. 2021. *Causation*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108588300.

Field, Hartry. 1978. "Mental Representation." *Erkenntnis* 13 (1): 9–61. http://doi.org/10.1007/BF00160888.

Field, Hartry. 2003. "Causation in a Physical World." In *The Oxford Handbook of Metaphysics*, edited by Michael J. Loux and Dean W. Zimmerman, 435–460. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199284221.003.0015.

Gallow, J. Dmitri. 2022. "The Metaphysics of Causation." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/fall2022/entries/causation-metaphysics/.

Gebharter, Alexander. 2017a. “Causal Exclusion Without Physical Completeness and No Overdetermination.” *Abstracta* 10: 3–14.

Gebharter, Alexander. 2017b. “Causal Exclusion and Causal Bayes Nets.” *Philosophy and Phenomenological Research* 95 (2): 353–375. https://onlinelibrary.wiley.com/doi/abs/10.1111/phpr.12247.

Gibbons, John. 2006. “Mental Causation Without Downward Causation.” *Philosophical Review* 115 (1): 79–103. http://doi.org/10.1215/00318108-115-1-79.

Goldman, Alvin I. 1969. “The Compatibility of Mechanism and Purpose.” *Philosophical Review* 78 (4): 468–482. https://doi.org/10.2307/2184199.

Goldman, Alvin I. 1970. *A Theory of Human Action*. Englewood Cliffs NJ: Prentice Hall.

Halpern, Joseph Y. 2016. “Appropriate Causal Models and the Stability of Causation.” *The Review of Symbolic Logic* 9 (1): 76–102. http://doi.org/10.1017/S1755020315000246.

Halpern, Joseph Y., and Christopher Hitchcock. 2010. “Actual Causation and the Art of Modeling.” In *Heuristics, Probability and Causality: A Tribute to Judea Pearl,* edited by Rina Dechter, Hector Geffner, and Joseph Y. Halpern, 383–406. London: College Publications.

Halpern, Joseph Y., and Judea Pearl. 2005. “Causes and Explanations: A Structural-Model Approach. Part i: Causes.” *British Journal for the Philosophy of Science* 56 (4): 843–887. http://doi.org/10.1093/bjps/axi147.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge MA: MIT Press.

Hausman, Daniel M. 2005. “Causal Relata: Tokens, Types, or Variables?” *Erkenntnis* 63 (1): 33–54. http://doi.org/10.1007/s10670-005-0562-6.

Heckman, James J. 2005. “The Scientific Model of Causality.” *Sociological Methodology* 35 (1): 1–98. http://doi.org/10.1111/j.0081-1750.2006.00163.x.

Henderson, David K. 1994. “Accounting for Macro-Level Causation.” *Synthese* 101 (2): 129–156. https://doi.org/10.1007/BF01064014.

Hendry, Robin. 2006. “Is There Downward Causation in Chemistry?” In *Philosophy of Chemistry: Synthesis of a New Discipline*, edited by Davis Baird, Eric Scerri, and Lee McIntyre, 242: 173–189. Boston Studies in the Philosophy of Science. Dordrecht: Springer. https://doi.org/10.1007/1-4020-3261-7_9.

Hendry, Robin. 2010a. “Emergence Vs. Reduction in Chemistry.” In *Emergence in Mind*, edited by Cynthia Macdonald and Graham Macdonald, 205–21. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199583621.003.0014.

Hendry, Robin. 2010b. “Ontological Reduction and Molecular Structure.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 41 (2): 183–191. https://doi.org/10.1016/j.shpsb.2010.03.005.

Hendry, Robin. 2017. “Prospects for Strong Emergence in Chemistry.” In *Philosophical and Scientific Perspectives on Downward Causation*, edited by Michele Paolini Paoletti and Francesco Orilia, 146–163. New York: Routledge. https://doi.org/10.4324/9781315638577-9.

Hitchcock, Christopher. 2001. “The Intransitivity of Causation Revealed in Equations and Graphs.” *Journal of Philosophy* 98 (6): 273–299. https://doi.org/10.2307/2678432.

Hitchcock, Christopher. 2004. “Routes, Processes, and Chance-Lowering Causes.” In *Cause and Chance: Causation in an Indeterministic World*, edited by Phil Dowe and Paul Noordhof, 138–152. London: Routledge. https://doi.org/10.4324/9780203494660.

Hitchcock, Christopher. 2007. “Prevention, Preemption, and the Principle of Sufficient Reason.” *Philosophical Review* 116 (4): 495–532. http://doi.org/10.1215/00318108-2007-012.

Hitchcock, Christopher. 2009. “Causal Modelling.” In *The Oxford Handbook of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Peter Menzies, 299–314. Oxford: Oxford University Press. http://doi.org/10.1093/oxfordhb/9780199279739.003.0015.

Hitchcock, Christopher. 2012. “Events and Times: A Case Study in Means-Ends Metaphysics.” *Philosophical Studies* 160 (1): 79–96. http://doi.org/10.1007/s11098-012-9909-4.

Hitchcock, Christopher. 2023. “Causal Models.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/spr2023/entries/causal-models/.

Hoffman-Kolss, Vera. 2014. "Interventionism and Higher-Level Causation." *International Studies in the Philosophy of Science* 28 (1): 49–64. http://doi.org/10.1080/02698595.2014.915653.

Horgan, Terence. 1997. "Kim on Mental Causation and Causal Exclusion." *Noûs* 31: 165–184. https://doi.org/10.1111/0029-4624.31.s11.8.

Humphreys, Paul W. 1989. *The Chances of Explanation*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400860760.

Ismael, Jenann. 2009. "Probability in Deterministic Physics." *Journal of Philosophy* 106 (2): 89–108. https://doi.org/10.5840/jphil2009106214.

Ismael, Jenann. 2011. "A Modest Proposal about Chance." *Journal of Philosophy* 108 (8): 416–442. https://doi.org/10.5840/jphil2011108822.

Jackson, Frank, and Philip Pettit. 1988. "Functionalism and Broad Content." *Mind* 97 (387): 381–400. http://doi.org/10.1093/mind/XCVII.387.381.

Jackson, Frank, and Philip Pettit. 1990a. "Causation in the Philosophy of Mind." *Philosophy and Phenomenological Research* 50: 195–214. http://doi.org/10.2307/2108039.

Jackson, Frank, and Philip Pettit. 1990b. "Program Explanation: A General Perspective." *Analysis* 50 (2): 107–117. http://doi.org/10.2307/3328853.

Kaiserman, Alex. 2020. "Interventionism and Mental Surgery." *Erkenntnis* 85 (4): 919–935. https://doi.org/10.1007/s10670-018-0059-8.

Kendler, Kenneth S., and John Campbell. 2009. "Interventionist Causal Models in Psychiatry: Repositioning the Mind-Body Problem." *Psychological Medicine* 39 (6): 881–887. http://doi.org/10.1017/S0033291708004467.

Kim, Jaegwon. 1973. "Causes and Counterfactuals." *Journal of Philosophy* 70 (17): 570–572. https://doi.org/10.2307/2025312.

Kim, Jaegwon. 1989. "Mechanism, Purpose, and Explanatory Exclusion." *Philosophical Perspectives* 3: 77–108. http://doi.org/10.2307/2214264.

Kim, Jaegwon. 1997. "Does the Problem of Mental Causation Generalize?" *Proceedings of the Aristotelian Society* 97 (3): 281–297. http://doi.org/10.1111/1467-9264.00017.

Kim, Jaegwon. 1998. *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation*. Cambridge MA: MIT Press. https://doi.org/10.7551/mitpress/4629.001.0001.

Kim, Jaegwon. 2003. "Blocking Causal Drainage and Other Maintenance Chores with Mental Causation." *Philosophy and Phenomenological Research* 67 (1): 151–176. http://doi.org/10.1111/j.1933-1592.2003.tb00030.x.

Kim, Jaegwon. 2005. *Physicalism, or Something Near Enough*. Princeton: Princeton University Press. https://doi.org/10.1515/9781400840847.

Korb, Kevin B., Lucas R. Hope, Ann E. Nicholson, and Karl Axnick. 2004. "Varieties of Causal Intervention." In *PRICAI 2004: Trends in Artificial Intelligence*, edited by Chengqi Zhang, Hans W. Guesgen, and Wai-Kiang Yeap, 322–331. Berlin: Springer. https://doi.org/10.1007/978-3-540-28633-2_35.

List, Christian, and Peter Menzies. 2009. "Nonreductive Physicalism and the Limits of the Exclusion Principle." *Journal of Philosophy* 106 (9): 475–502. http://doi.org/10.2307/20620197.

Loewer, Barry. 2008. "Why There Is Anything Except Physics." In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 149–163. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.001.0001.

Loewer, Barry. 2009. "Why Is There Anything Except Physics?" *Synthese* 170 (2): 217–233. http://doi.org/10.1007/s11229-009-9580-2.

Loewer, Barry. 2015. "Mental Causation: The Free Lunch." In *Qualia and Mental Causation in a Physical World: Themes from the Philosophy of Jaegwon Kim*, edited by David Sosa, Terence Horgan, and Marcelo Sabatés, 40–63. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139939539.004.

Lowe, E. J. 2000. "Causal Closure Principles and Emergentism." *Philosophy* 75 (294): 571–585. https://doi.org/10.1017/S003181910000067X.

Lowe, E. J. 2003. “Physical Causal Closure and the Invisibility of Mental Causation.” In *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, edited by Sven Walter and Heinz-Dieter Heckmann, 137–154. Exeter: Imprint Academic.

Mackie, John L. 1974. *The Cement of the Universe: A Study of Causation*. Oxford: Oxford University Press. http://doi.org/10.1093/0198246420.001.0001.

Malcolm, Norman. 1968. “The Conceivability of Mechanism.” *Philosophical Review* 77 (1): 45–72. https://doi.org/10.2307/2183182.

Markowetz, Florian, Steffen Grossmann, and Rainer Spang. 2005. “Probabilistic Soft Interventions in Conditional Gaussian Networks.” In *IN 10TH AI/STATS*, edited by Robert Cowell and Zoubin Ghahramani, 214–221. Savannah Hotel, Barbados: The Society for Artificial Intelligence and Statistics. https://www.gatsby.ucl.ac.uk/aistats/fullpapers/139.pdf.

Marras, Ausonio. 1998. “Kim’s Principle of Explanatory Exclusion.” *Australasian Journal of Philosophy* 76 (3): 439–451. https://doi.org/10.1080/00048409812348551.

McDermott, Michael. 1995. “Redundant Causation.” *British Journal for the Philosophy of Science* 46 (4): 523–544. http://doi.org/10.1093/bjps/46.4.52.

McDermott, Michael. 2002. “Causation: Influence Versus Sufficiency.” *Journal of Philosophy* 99 (2): 84–101. https://doi.org/10.5840/jphil200299219.

McDonald, Jenn. forthcoming. “Essential Structure for Causal Models.” *Australasian Journal of Philosophy*, forthcoming.

McDonald, Jenn. 2023. “Causal Models and Contrastivism.”, unpublished manuscript.

Mellor, D. H. 1995. *The Facts of Causation*. London: Routledge.

Melnyk, Andrew. 2003. *A Physicalist Manifesto: Thoroughly Modern Materialism*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511498817.

Menzies, Peter. 2003. “The Causal Efficacy of Mental States.” In *Physicalism and Mental Causation: The Metaphysics of Mind and Action*, edited by Sven Walter and Heinz-Dieter Heckmann, 195–224. Exeter: Imprint Academic.

Menzies, Peter. 2010. “Reasons and Causes Revisited.” In *Naturalism and Normativity*, edited by Mario De Caro and David Macarthur, 142–170. New York: Columbia University Press.

Menzies, Peter, and Christian List. 2010. “The Causal Autonomy of the Special Sciences.” In *Emergence in Mind*, edited by Cynthia Macdonald and Graham Macdonald, 108–128. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199583621.003.0008.

Noordhof, Paul. 1997. “Making the Change: The Functionalist’s Way.” *British Journal for the Philosophy of Science* 48 (2): 233–250. http://doi.org/10.1093/bjps/48.2.233.

Papineau, David. 2001. “The Rise of Physicalism.” In *Physicalism and Its Discontents*, edited by Carl Gillett and Barry Loewer, 3–36. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511570797.002.

Papineau, David. 2002. *Thinking about Consciousness*. Oxford: Oxford University Press. http://doi.org/10.1093/0199243824.001.0001.

Patterson, Sarah. 2005. “Epiphenomenalism and Occasionalism: Problems of Mental Causation, Old and New.” *History of Philosophy Quarterly* 22 (3): 239–257.

Peacocke, Christopher. 1993a. “Externalist Explanation.” *Proceedings of the Aristotelian Society* 93 (3): 203–230. https://doi.org/10.1093/aristotelian/93.1.203.

Peacocke, Christopher. 1993b. “Review of the Imagery Debate.” *Philosophy of Science* 60 (4): 675–677. https://doi.org/10.1086/289774.

Pearl, Judea. 2000. *Causality*. Cambridge: Cambridge University Press.

Pereboom, Derk, and Hilary Kornblith. 1991. “The Metaphysics of Irreducibility.” *Philosophical Studies* 63 (2): 125–145. http://doi.org/10.1007/BF00381684.

Polger, Thomas W., Lawrence A. Shapiro, and Reuben Stern. 2018. “In Defense of Interventionist Solutions to Exclusion.” *Studies in History and Philosophy of Science Part* A 68 (April): 51–57. https://doi.org/10.1016/j.shpsa.2018.01.012.

Price, Huw, and Richard Corry. 2007. *Causation, Physics and the Constitution of Reality: Russell’s Republic Revisited*. Edited by Huw Price and Richard Corry. Oxford: Oxford University Press.

Putnam, Hilary. 1975. “Philosophy and Our Mental Life.” In *Mind, Language and Reality*, 2: 291–303. Philosophical Papers. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511625251.016.

Raatikainen, Panu. 2010. “Causation, Exclusion, and the Special Sciences.” *Erkenntnis* 73 (3): 349–363. https://doi.org/10.1007/s10670-010-9236-0.

Rescorla, Michael. 2012. “Are Computational Transitions Sensitive to Semantics?” *Australasian Journal of Philosophy* 90 (4): 703–721. https://doi.org/10.1080/00048402.2011.615333.

Rescorla, Michael. 2014. “The Causal Relevance of Content to Computation.” *Philosophy and Phenomenological Research* 88 (1): 173–208. http://doi.org/10.1111/j.1933-1592.2012.00619.x.

Rescorla, Michael. 2018. “An Interventionist Approach to Psychological Explanation.” *Synthese* 195 (5): 1909–1940. https://doi.org/10.1007/s11229-017-1553-2.

Robb, David, John Heil, and Sophie Gibb. 2023. “Mental Causation.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/spr2023/entries/mental-causation/.

Russell, Bertrand. 1913. “On the Notion of Cause.” *Proceedings of the Aristotelian Society* 13 (1): 1–26. https://doi.org/10.1093/aristotelian/13.1.1.

Schaffer, Jonathan. 2016. “The Metaphysics of Causation.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford: Stanford University. https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/.

Schlosser, Markus E. 2009. “Non-Reductive Physicalism, Mental Causation, and the Nature of Action.” In *Reduction: Between the Mind and the Brain*, edited by Alexander Hieke and Hannes Leitgeb, 73–90. Frankfurt: Ontos Verlag. https://doi.org/10.1515/9783110328851.73.

Segal, Gabriel. 2004. “Reference, Causal Powers, Externalist Intuitions, and Unicorns.” In *The Externalist Challenge*, edited by Richard Schantz, 329–344. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110915273.329.

Segal, Gabriel. 2009. “The Causal Inefficacy of Content.” *Mind and Language* 24 (1): 80–102. http://doi.org/10.1111/j.1468-0017.2008.01354.x.

Segal, Gabriel, and Elliott Sober. 1991. “The Causal Efficacy of Content.” *Philosophical Studies* 63 (1): 1–30. http://doi.org/10.1007/BF00375995.

Shapiro, Lawrence A. 2010. “Lessons from Causal Exclusion.” *Philosophy and Phenomenological Research* 81 (3): 594–604. http://doi.org/10.1111/j.1933-1592.2010.00382.x.

Shapiro, Lawrence A., and Elliott Sober. 2007. “Epiphenomenalism: The Dos and the Don’ts.” In *Thinking about Causes: From Greek Philosophy to Modern Physics*, edited by Peter Machamer and Gereon Wolters, 235–264. Pittsburgh-Konstanz Series in the Philosophy and History of Science. Pittsburgh: University of Pittsburgh Press.

Sklar, Lawrence. 2003. “Dappled Theories in a Uniform World.” *Philosophy of Science* 70 (2): 424–441. http://doi.org/10.1086/375476.

Sober, Elliott. 1999a. “Physicalism from a Probabilistic Point of View.” *Philosophical Studies* 95 (1–2): 135–174. http://doi.org/10.1023/A:1004519608950.

Sober, Elliott. 1999b. “The Multiple Realizability Argument Against Reductionism.” *Philosophy of Science* 66 (4): 542–564. https://doi.org/10.1086/392754.

Statham, Georgie. 2018. “Woodward and Variable Relativity.” *Philosophical Studies* 175 (4): 885–902. https://doi.org/10.1007/s11098-017-0897-2.

Stern, Reuben, and Benjamin Eva. 2023. “Antireductionist Interventionism.” *British Journal for the Philosophy of Science* 74 (1): 241–267. https://doi.org/10.1086/714792.

Stich, Stephen. 1983. *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge MA: MIT Press.

Strevens, Michael. 2007. “Review of Woodward, Making Things Happen.” *Philosophy and Phenomenological Research* 74 (1): 233–249. http://doi.org/10.1111/j.1933-1592.2007.00012.x.

Thomasson, Amie. 1998. “A Nonreductivist Solution to Mental Causation.” *Philosophical Studies* 89 (2–3): 181–195. http://doi.org/10.1023/A:1004280812099.

Tye, Michael. 1991. *The Imagery Debate*. Cambridge MA: MIT Press.

Weslake, Brad. unpublished. “A Partial Theory of Actual Causation.”

Weslake, Brad. 2010. "Explanatory Depth." *Philosophy of Science* 77 (2): 273–294. http://doi.org/10.1086/651316.

Weslake, Brad. 2017. "Difference-Making, Closure and Exclusion." In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 215–231. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198746911.003.0011.

Woodward, James. 2003. *Making Things Happen: A Theory of Causal Explanation*. New York: Oxford University Press. http://doi.org/10.1093/0195155270.001.0001.

Woodward, James. 2008a. "Mental Causation and Neural Mechanisms." In *Being Reduced: New Essays on Reduction, Explanation and Causation*, edited by Jesper Kallestrup and Jakob Hohwy, 218–262. Oxford: Oxford University Press. http://doi.org/10.1093/acprof:oso/9780199211531.003.0013.

Woodward, James. 2008b. "Response to Strevens." *Philosophy and Phenomenological Research* 77 (1): 193–212. http://doi.org/10.1111/j.1933-1592.2008.00181.x.

Woodward, James. 2015a. "Interventionism and Causal Exclusion." *Philosophy and Phenomenological Research* 91 (2): 303–347. http://doi.org/10.1111/phpr.12095.

Woodward, James. 2015b. "Methodology, Ontology, and Interventionism." *Synthese* 192 (11): 3577–3599. https://doi.org/10.1007/s11229-014-0479-1.

Woodward, James. 2016. "The Problem of Variable Choice." *Synthese* 193 (4): 1047–1072. https://doi.org/10.1007/s11229-015-0810-5.

Woodward, James. 2017. "Intervening in the Exclusion Argument." In *Making a Difference: Essays on the Philosophy of Causation*, edited by Helen Beebee, Christopher Hitchcock, and Huw Price, 251–268. Oxford: Oxford University Press. http://doi.org/10.1093/oso/9780198746911.003.0013.

Woodward, James. 2020. "Causal Complexity, Conditional Independence, and Downward Causation." *Philosophy of Science* 87 (5): 857–867. https://doi.org/10.1086/710631.

Woodward, James. 2021a. "Downward Causation and Levels." In *Levels of Organization in the Biological Sciences*, edited by Daniel S. Brooks, James DiFrisco, and William C. Wimsatt, 175–194. Cambridge MA: MIT Press. https://doi.org/10.7551/mitpress/12389.003.0013.

Woodward, James. 2021b. "Downward Causation Defended." In *Top-Down Causation and Emergence*, edited by Jan Voosholz and Markus Gabriel, 217–52. Cham: Springer. https://doi.org/10.1007/978-3-030-71899-2_9.

Woodward, James. 2021c. "Explanatory Autonomy: The Role of Proportionality, Stability, and Conditional Irrelevance." *Synthese* 198 (1): 237–65. https://doi.org/10.1007/s11229-018-01998-6.

Woodward, James. 2022. "Levels, Kinds and Multiple Realizability: The Importance of What Does Not Matter." In *Levels of Reality in Science and Philosophy*, edited by Stavros Ioannidis, Gal Vishne, Meir Hemmo, and Orly Shenker, 261–292. Cham: Springer. https://doi.org/10.1007/978-3-030-99425-9_14.

Worley, Sara. 1993. "Mental Causation and Explanatory Exclusion." *Erkenntnis* 39 (3): 333–358. https://doi.org/10.1007/BF01128507.

Yablo, Stephen. 1992. "Mental Causation." *Philosophical Review* 101 (2): 245–280. https://doi.org/10.2307/2185535.

Yablo, Stephen. 1997. "Wide Causation." *Noûs* 31: 251–281. https://doi.org/10.1111/0029-4624.31.s11.12.

5
Interventionist Causal Exclusion and the Challenge of Mixed Models

Vera Hoffmann-Kolss

1 Introduction

Higher-level causal statements are ubiquitous. They occur in everyday contexts as well as in scientific practice. Many claims of the so-called *special sciences*, such as biology, neurophysiology, or the social sciences, seem to presuppose that it is possible to empirically discover higher-level causal relations and that knowledge about higher-level causal relations plays a crucial role in our best empirical theories of the world. In the past twenty years, many philosophers have argued that Woodward's interventionist theory of causation and causal modelling approaches in general are particularly well suited to describe causal relations holding at multiple levels. It has turned out, however, that interventionism faces a difficulty when dealing with higher-level causal claims.

In a nutshell, the problem is this: according to the interventionist account of causation, cause-effect relations are characterized as relations between variables whose values represent events, states, or properties. If variables X and Y occur in a causal model M, then the interventionist criterion implies that X is causally relevant to Y iff there is a possible intervention on the value of X which changes the value or the probability distribution of Y, provided that the values of all other variables in M that are not on a causal path between X and Y are kept fixed (Hitchcock 2001, 2007; Spirtes, Glymour, and Scheines 2000; Woodward 2003). The latter condition, that the values of variables not on a causal path between X and Y have to be kept fixed, is supposed to exclude possible confounding factors that might lead to false conclusions about the relationship between X and Y. However, in the context of higher-level causal claims, it creates a now well-known problem that has been dubbed the "interventionist causal exclusion problem."

A widely accepted minimal condition for the relationship between properties occurring at different levels is that they satisfy a supervenience constraint: for each higher-level property H, there is a set of lower-level properties, such that

Vera Hoffmann-Kolss, *Interventionist Causal Exclusion and the Challenge of Mixed Models* In: *Levels of Explanation*.
Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Vera Hoffmann-Kolss 2024.
DOI: 10.1093/oso/9780192862945.003.0006

it is impossible to change H without changing at least some of the lower-level properties. This immediately raises the question of whether the lower-level properties in the supervenience base are a confounder that has to be kept fixed when determining whether H is causally relevant to some other property. If they were, interventions on higher-level properties would be impossible, and all higher-level properties would be causally excluded by the lower-level properties upon which they supervene and come out causally inert. However, such a radical consequence would undermine one of the central advantages of interventionism, that is, the possibility to describe causal relations in the special sciences.

The interventionist causal exclusion problem occurs especially if models can be mixed or hybrid in the sense that they not only represent causal dependence relations, but both causal and non-causal dependencies, in particular, dependencies characterized by the supervenience relation. A number of authors seem to agree that even though such mixed models deviate from the orthodoxy of causal modelling (according to which the only dependence relations in causal models are supposed to be causal ones), they provide a fruitful approach to understanding complex causal networks, that is, networks spanning multiple levels (Kistler 2013; Kroedel 2020; Shapiro 2010; Stern and Eva 2023; Woodward 2015; for discussion, see Eronen and Brooks 2014). Obviously, this requires a solution to the interventionist version of the causal exclusion problem, and the standard approach is to simply make an exception for variables contained in a supervenience base of the cause variable under consideration: these variables need not be kept fixed when intervening on the supervenient variable (the *locus classicus* of this view is Woodward 2015).

The first aim of this paper is to show that the problems raised by mixed models run deeper than is usually assumed. I begin with an introduction to Woodward's interventionist theory of causation (Section 2) and explain why it leads to the interventionist version of the causal exclusion problem (Section 3). I then present the standard solution to this problem and argue that it must be rejected, since it leads to a trivialization of the condition that confounders must be held fixed (Section 4).

The second aim of the paper is to embed the question of how to deal with higher-level causal relations in the more general discussion about the question of how to select the variables to be included in a causal model. I discuss a metaphysical restriction on the variables that constitute causal models, according to which mixed models cannot be apt (Section 5), as well as two other types of metaphysical restriction that can be imposed on multilevel causal structures, one based on the notion of grounding, the other based on the notion of naturalness (Section 6). I conclude by pointing out that whichever option one chooses, a viable theory of multilevel causal structures rests on strong metaphysical commitments (Section 7).

2 Interventionism and causal models

As pointed out above, interventionism describes causal relations as relations between variables whose values stand for events, states, or properties.[1] Causal structures are represented by directed causal graphs consisting of two elements: (a) a set V of vertices constituted by variables standing in causal relations to each other and (b) a set of directed edges connecting those vertices. If a sequence of variables $\{X_1, \ldots X_n\}$ is such that for any i with $1 \leq i < n$ there is a directed edge from X_i to X_{i+1}, then the sequence is called a 'directed path leading from X_1 to X_n' (Spirtes, Glymour, and Scheines 2000; Woodward 2003). Directed paths represent causal relevance relations. If X and Y are variables occurring in a causal model $\mathcal{M}$, which is constituted by a variable set V, then X is classified as causally relevant to Y according to $\mathcal{M}$ iff there is a possible intervention on the value of X which changes the value or the probability distribution of Y, provided that the values of all other variables in V that are not on a causal path between X and Y are kept fixed at some value (Hitchcock 2001, 2007; Woodward 2003).[2]

To see how this theory of causation works, consider the causal relations holding between the following four variables:

H: 1 if person A has a headache; 0 otherwise
P: 1 if person A takes a painkiller; 0 otherwise
I: 1 if person A takes some ineffective drug; 0 otherwise
W: 1 if person A drinks a large glass of water; 0 otherwise

Suppose that there are no placebo effects and that P is causally relevant to H, whereas I is not causally relevant to H. Furthermore, assume that having a large glass of water has a causal influence on A's headache. Dehydration can cause headaches, and headaches often improve with water intake. Finally, assume that A's usual procedure for taking drugs (both ineffective ones and painkillers) is to swallow them with a large glass of water. Accordingly, P and W are both causally relevant to H, whereas I is not causally relevant to H. Moreover, W is a possible confounder for the relation between P and H and for the relation between I and H. The causal graph of this model is shown in Figure 5.1.

The interventionist criterion of causation correctly implies that there is a causal connection between P and H, but not between I and H. For instance, if

[1] For the sake of simplicity, I will assume that the values of variables can stand for events, states, *or* properties and ignore the distinction between type and token causation.

[2] Throughout the paper, I focus primarily on Woodward's interventionist theory of causation. However, the argument could in principle be adapted to other versions of the causal modelling approach, for instance, to structural equation approaches (Halpern and Pearl 2005; Hitchcock 2007) and causal Bayes nets theories (Spirtes, Glymour, and Scheines 2000). Note, however, that many proponents of these alternative approaches are skeptical of metaphysical solutions such as those suggested in the end of this paper, and tend toward more pragmatic solutions.

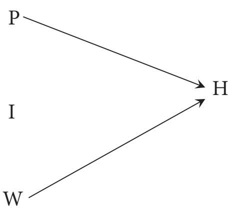

Figure 5.1 Causal graph showing the relationships between the variables P ("painkiller"), I ("ineffective drug"), W ("water consumption"), and H ("headache"). There is no causal relation between I and H, and W is a possible confounding factor for the causal relation between P and H.

person A has a headache, and takes a painkiller, this will increase the probability that her headache will disappear, regardless of the values of I and W. On the other hand, if person A does not take a painkiller, but drinks a large glass of water, then taking the ineffective drug will not increase the probability that her headache will disappear. More technically, if P = 0 and W = 0 and if P and W are kept fixed at those values, then changing the value of I from I = 0 to I = 1 by an intervention does not change the probability distribution of H. Since an analogous consideration holds for the other three possible combinations of values of P and W, there is no combination of values of P and W, such that if P and W were kept fixed at those values, then changing the value of I would change the probability distribution of H.

Moreover, not only does the model correctly imply that P is causally relevant to H, whereas I is not, it also provides information about possible confounders by showing that W is one of the variables that must be held fixed when determining whether P and I are causes of H. In general, it seems to be an advantage of a model if it can provide this kind of information. As Woodward puts it:

> [T]he bare claim that X causes Y is not very informative. From the perspective of a manipulability account, what one would really like to know is not just whether there is some manipulation of (or intervention on) X that will change Y; that is, whether it is true that X causes Y. One would also like to have more detailed information about just which interventions on X will change Y (and in what circumstances) and how they will change Y. (Woodward 2003, 66, see also 68–70; Statham 2018)

This may suggest that models containing more variables are generally preferable to models containing fewer variables. Models containing more variables tend to provide more information about possible confounders and the background conditions in which causal relations hold. The crucial question is whether this also implies that mixed models are preferable to non-mixed models, since mixed models tend to include more variables, and thus to provide more information. I will return to this question below.

3 Interventionist causal exclusion

The causal structure leading to the classical causal exclusion argument à la Kim consists of three variables: a variable M representing some mental property, and two variables, P and P*, representing physical properties. It is assumed that P is causally relevant to P* and that M supervenes on P. The worry raised by the classical version of the causal exclusion argument is that if M is also causally relevant to P*, P* will be overdetermined by P and M, and since this type of overdetermination is considered problematic, it is concluded that all the causal work is done by P and there is no independent causal role left for M.

The interventionist version of the causal exclusion argument, as described in several of Baumgartner's papers, does not rely on overdetermination worries, but follows directly from the structure of the causal exclusion scenario and the definition of interventionist causation. Consider a mixed model constituted by the variable set V = {M, P, P*}. The interventionist criterion of causation has the form of an existence condition. If M is to be causally relevant to P* according to the model constituted by V, there must be an intervention on the value of M that keeps all variables in V that are not on a causal path between M and P* fixed at some value. This intervention must be possible, but needs not be actually carried out. Given that P is contained in V and not located on a causal path between M and P*, P is one of the variables that must be kept fixed when carrying out an intervention on M with respect to P*. However, since M supervenes on P, it is not possible to intervene on the value of M without changing the value of P.[3] Therefore, no intervention on M satisfies the interventionist criterion of causation (relative to V), and M comes out causally inert (Baumgartner 2009, 2010).

Baumgartner does not explicitly define the notion of supervenience used in this argument. He merely points out that it is a non-causal dependence relation holding between sets of properties, so that all changes in the supervening properties are necessarily accompanied by a change in some property in the supervenience base (Baumgartner 2009, 2010). However, it seems to be in accordance with the general

[3] Woodward discusses how the notion of possibility used here should be understood, pointing out that "the reference to 'possible' interventions … does not mean 'physically possible'; instead, an intervention on X with respect to Y will be 'possible' as long as it is logically or conceptually possible for a process meeting the conditions for an intervention on X with respect to Y to occur" (Woodward 2003, 132). The question of how to interpret the notion of possibility in this context deserves more attention than I can give it here. In what follows, however, I will assume that if the mental supervenes on the physical, then it will be impossible to manipulate the mental without the physical in a sense of "impossible" that excludes such interventions.

idea of the interventionist exclusion problem to adapt the classical notion of strong supervenience to the interventionist framework in the following way:

> **Variable supervenience**: A variable X supervenes on a set of variables $\{Z_1, \ldots , Z_n\}$ iff necessarily, for any value of X, x, if X = x, then there is a combination of the values of $Z_1, \ldots , Z_n$, such that $Z_1, \ldots Z_n$ assume these values and, necessarily, whenever $Z_1, \ldots , Z_n$ assume these values, then X = x.

There are two things to note about this definition. First, it is compatible with the observation that the supervenience base or some variable can consist of one or of several variables. In causal exclusion contexts, some authors assume that M supervenes on a single physical variable whose set of values exhausts the set of all possible realizers of the values of M (e.g., Eronen and Brooks 2014; Shapiro and Sober 2007). Other authors assume that the supervenience base of M is a set of binary variables, each of which represents whether or not a certain realizer property is instantiated (e.g., Zhong 2020). The definition of supervenience just proposed allows me to remain neutral concerning this question.

The second thing to note about the definition of variable supervenience is that it can be satisfied by many sets of variables that stand in purely formal relations to each other. For instance, if Y = f(X), that is, Y is a (mathematical) function of X, then Y supervenes on {X}, according to the definition given above. Moreover, the supervenience definition is not restricted to the mental-physical case. Many physicalists assume that non-mental higher-level properties, such as biological or chemical properties, supervene on physical properties, and such relations are also covered by the definition of variable supervenience.

It follows that the interventionist exclusion argument generalizes (as does the classical version of the exclusion argument, see, e.g., Bontly 2002). Whenever a model contains variables standing in a supervenience relation to each other, it is impossible to intervene on the supervenient variable and keep the values in the supervenience base fixed. Therefore, all variables supervening on some other set of variables contained in the same causal model will be deprived of all their causal powers.[4]

Worse, the problem also occurs in non-mixed models. This is because, according to Woodward's theory, whether X is causally relevant to Y depends on whether there is a possible *intervention* on X which changes the value or the probability distribution of Y. One of the conditions that an intervention must satisfy is that it must not be "correlated with any variable Z that is causally relevant to Y through a route

[4] I will leave open here whether Baumgartner's original version of the argument is intended to show that interventionism faces this more general difficulty. For even if Baumgartner's original argument refers only to the more specific case of the relationship between the mental and the physical, the more general version deserves attention as well.

that excludes X" (Woodward and Hitchcock 2003, 13). Z, the variable of which the intervention must be independent, need not be included in the same causal model as X and Y. Therefore, even if one considers a non-mixed model, containing only the variables M and P*, any possible intervention on M with respect to P* will be correlated with a change in P—another variable that is causally relevant to M via a route that excludes M. Thus, given that M supervenes on P, there is no possible intervention on M with respect to P*, and M turns out to be causally inert.[5]

The standard response to this problem is that something goes wrong when the condition that all off-path variables have to be kept fixed is applied to variables standing in a supervenience relation to the cause variable X under consideration: variables that are members of some supervenience base of X simply do not count as confounders that have to be controlled for. As Shapiro puts it:

> When investigating whether a supervening property is a cause, one must not ask whether the supervening property has causal influence in addition to the causal influence of its base. This question suggests the wrong kind of test, i.e. a test in which the base is held fixed while the supervening property is changed. (Shapiro 2010, 8)

For instance, consider the following binary variable:

Pr: 1 if the monthly precipitation in region r in the summer is higher than 70 millimeters; 0 otherwise

Pr could be included in a causal model describing the relationship between the amount of precipitation in region r and the growth of a certain plant. However, it would be pointless to require that the value of Pr has to be manipulated, whereas the value of another variable describing the exact amount of rain falling in region r in the summer has to be held fixed. Shapiro argues that manipulating M, while holding the value of P fixed (in the causal exclusion schema), would be equally pointless.

In a similar vein, Woodward proposes the following modified definition of the notion of an intervention:

> Put slightly differently, an intervention I on X with respect to Y will (a) fix the value of SB(X) [i.e., the supervenience base of X] in a way that respects the

[5] It is controversial among proponents of interventionism to what extent the interventionist criterion of causation is model-relative, and whether model-relativity would be problematic (McCain 2015; Rolffs 2023; Statham 2018; Strevens 2007, 2008; Woodward 2008b). I will not enter into that discussion, since all that is necessary for the present argument is the observation that the interventionist exclusion problem arises not only with respect to mixed models, but with respect to any model that includes variables that supervene on other variables, regardless of whether those other variables are also included in the model.

> supervenience relationship between X and SB(X), and (b) the requirements in the definition (IV) [i.e., the definition of an intervention] are understood as applying only to those variables that are causally related to X and Y or are correlated with them but not to those variables that are related to X and Y as a result of supervenience relations or relations of definitional dependence. (Woodward 2015, 334)

Accordingly, Woodward's proposal is to relax the criterion of interventionist causation in such a way that not all variables not on a causal path between the variables under consideration must be held at their actual values, but only those variables that do not stand in a supervenience relation to the cause variable. *Prima facie*, this is a simple and effective strategy to stick with the idea of mixed models, including models describing the classical causal exclusion schema, and to avoid the interventionist exclusion problem at the same time. I will argue in the next section, however, that this strategy involves a hitherto unnoticed difficulty.[6]

4 The problem of trivialization

Here is a first-pass formulation of the interventionist criterion of causation that contains an exemption clause for variables that are in the supervenience base of the cause variable under consideration:

> **Interventionist causation with exemption clause**: X is classified as causally relevant to Y iff there is an intervention on the value of X which changes the value or the probability distribution of Y, provided that the values of all other

[6] In the debate on the interventionist exclusion problem, there are three further strands of discussion which I cannot address in this paper. The first relies on the consideration that supervenience relations between variables are formally similar to causal relations between variables and that this might lead to a further version of the interventionist exclusion problem (Gebharter 2017; for discussion, see Stern and Eva 2023). I have argued elsewhere that metaphysical dependence relations, such as the supervenience relation, have to be clearly distinguished from causal relations (Hoffmann-Kolss 2022), and I will not get further into this debate here.

The second strand of discussion revolves around the question of whether interventionism solves the classical causal exclusion problem (rather than creating the particularly detrimental interventionist version of it that is the topic of this paper). Several authors have argued that interventionism is particularly well suited to solve the classical causal exclusion problem (e.g., List and Menzies 2009; Menzies 2008; Raatikainen 2010; Woodward 2008a). I have argued against these approaches in a different paper (Hoffmann-Kolss 2014) and will not comment on them in the present context.

A third strand of discussion results from a recent observation by Blanchard that the standard strategy of avoiding the interventionist exclusion problem sometimes implies that too many higher-level dependencies are misclassified as causal, in particular that there are cases where properties of composite objects are classified as causally relevant to a certain effect, although it is intuitively clear that only the properties of one of the composite's parts are relevant to that effect (Blanchard 2023). Exploring the consequences of this problem will require future work.

> variables that are not on a causal path between X and Y and are not contained in some supervenience base of X are kept fixed at some value.

Note that this definition exempts both variables that are in the same model as X and Y and variables that are not in the same model as X and Y from the *has-to-be-held-fixed* condition, provided that they are in the supervenience base of X. However, it can be easily trivialized. To see this, note, first, that X supervenes on {X}, according to the above definition of supervenience. Furthermore, note that the supervenience relation is conserved under the addition of new elements to the supervenience base, that is, if X supervenes on a set of variables S, then X supervenes on every superset of S. It follows that X supervenes on every set of the form {X, Z}, where Z is some arbitrary variable. But then, every variable is contained in a supervenience base of X, and the exemption clause implies that *no* variable has to be kept fixed when the value of X is manipulated by an intervention. This in turn implies that no confounding factors have to be kept fixed. But that consequence would call the adequacy of the whole interventionist approach into question.

It seems that this result can be avoided if the exemption clause is interpreted differently. According to a more charitable interpretation, it should only imply that variables that are *really* related to the cause variable X under consideration need not be held fixed. A natural way to interpret this requirement is to say that the variables falling under the exemption clause should be those that are *non-redundant* elements of some supervenience base of X, and that Z is a non-redundant element of some supervenience base S of X iff X supervenes on S, Z∈S, and X does not supervene on S\{Z}. Accordingly, the interventionist criterion of causation with exemption clause can be modified as follows:

> **Interventionist causation with exemption clause***: X is classified as causally relevant to Y iff there is an intervention on the value of X which changes the value or the probability distribution of Y, provided that the values of all other variables that are not on a causal path between X and Y and are not non-redundant elements of some supervenience base of X are kept fixed at some value.

At first sight, this criterion fulfils the function we are looking for. In the classical causal exclusion schema, P is a non-redundant element of some supervenience base of M. Therefore, P falls under the exemption clause and does not have to be kept fixed when the value of M is changed by an intervention. To see whether the new criterion can also tackle confounding factors, reconsider the headache pill model introduced in Section 2 (the variables are listed again for the sake of convenience):

H: 1 if person A has a headache; 0 otherwise
P: 1 if person A takes a painkiller; 0 otherwise

I: 1 if person A takes some ineffective drug; 0 otherwise
W: 1 if person A drinks a large glass of water; 0 otherwise

Since W is a possible confounder of both the relationship between P and H and the relationship between I and H, W had better not fall under the exemption clause. It should be uncontroversial that P does not supervene on {W}, and I does not supervene on {W} either. *Prima facie*, this implies that W does not fall under the exemption clause and therefore has to be kept fixed when the value of P or the value of I is changed by an intervention. However, this conclusion is too rash. According to the criterion specified above, all variables that are a non-redundant element of *some* supervenience base of X fall under the exemption clause. This is crucial, since most variables have multiple supervenience bases. The classical causal exclusion schema, according to which M has a single variable P as its supervenience base, can only be considered a simplification of the multilevel structure really characterizing the relationship between mental states and the lower-level states underlying them. It is plausible to assume, for instance, that depending on the level of specification one is interested in, the supervenience base of M consists either of variables describing neuronal states or of variables describing microphysical states or of a combination of both. All these sets can provide a complete supervenience base of M, and if they do, their elements fall under the exemption clause.

However, if all variables that are a non-redundant element of *some* supervenience base of the cause variable fall under the exemption clause, a more complex version of the trivialization problem arises. To see this, consider a logically complex variable defined as follows:

I↔W: 1 if variables I and W have the same value; 0 otherwise

The set {I↔W} is not a supervenience base of I, since neither of the two possible values of I↔W necessitates a particular value of I. The set {I↔W, W}, by contrast, is a supervenience base of I. To see this, suppose that I = 1. Then there is a combination of values of I↔W and I, for instance, the combination I↔W = 1 and W = 1, that necessitates I = 1. And analogous reasoning holds if I = 0. Accordingly, W is a non-redundant element of some supervenience base of I. But then the exemption clause implies that W does not have to be kept fixed by an intervention on I with respect to H. This is problematic given that W is a confounder for the relationship between I and H.

Once the structure of this argument is clear, it can be easily generalized. Whenever a model contains two independent binary variables, X and Y, one can argue that Y need not be kept fixed by an intervention on X, since {X↔Y, Y} is a supervenience base of X containing Y as a non-redundant element. Analogously, if a model contains two independent variables, X and Y, that take real numbers as values (for instance, variables measuring the length, volume, mass, or temperature of something), then Y need not be kept fixed by an intervention on X, since {X+Y, Y} is a supervenience base of X, whereas {X+Y} is not, and this implies that Y is a non-redundant element of some supervenience base of X. It is plausible to assume that structurally similar arguments can be constructed for almost every pair of variables X and Y. But then the crucial condition that all possible confounding factors for the causal relations in a model have to be held fixed is trivialized, since almost all variables will be exempted from it.

An immediate response to this line of reasoning is that variables such as I↔W or X+Y are artificially created by logical or mathematical operations applied to variables already contained in a model, and that such variables should be excluded by imposing further constraints on the variables forming the relevant supervenience bases. However, introducing such purely formal constraints often leads to further complications that have to be met with further modifications and constraints, and in many cases this process does not make a criterion more informative. For instance, imposing a ban on variables that are logical or mathematical compounds of variables already contained in the model under consideration will not solve the problem, since one can create more complicated counterexamples evading this condition.

To see this, suppose that a model contains two unrelated variables, X and Y, and that $X = A_1 + A_2$ and $Y = A_3 + A_4$ (where $A_1 - A_4$ are unrelated variables assuming real numbers as values). $A_1 - A_4$ are not contained in the model, since they are only mathematical components of X and Y. Define Z_1 as the sum of A_1 and A_3, and Z_2 as the sum of A_2 and A_4, that is, $Z_1 = A_1 + A_3$ and $Z_2 = A_2 + A_4$. Then, $\{Z_1, Z_2, Y\}$ is a supervenience base of X, whereas $\{Z_1, Z_2\}$ is not.[7] Therefore, Y is a non-redundant element of some supervenience base of X and would not have to be held fixed by an intervention on X. It is plausible to assume that adding yet another formal restriction, for instance, excluding variables that are compounds of the *components* of variables already contained in the model, would just lead to the construction of more complicated counterexamples.

Such problems are common when purely formal criteria are used to define metaphysically substantial notions or ideas. There is an intuitive distinction between variables that are really related and variables that are just related by a logical or formal trick (as in the above examples). This distinction is metaphysically substantial, and it is not surprising that a purely formal criterion will fall short of characterizing it in an adequate way. A common move to solve this type of difficulty is to introduce metaphysically "thicker" notions, such as naturalness, grounding, or fundamentality. In the case at hand, one might claim, for instance, that the variables in the supervenience bases must not be less fundamental or less natural than

[7] $\{Z_1, Z_2, Y\}$ is a supervenience base of X, since $X = Z_1 + Z_2 - Y$. However, the values of Z_1 and Z_2 alone are not sufficient to determine the value of X, which is why $\{Z_1, Z_2\}$ is not a supervenience base of X.

the cause and effect variables under consideration (Hoffmann-Kolss 2022). I will briefly come back to this option at the end of the paper.

5 A ban on mixed models?

The interventionist causal exclusion problem arises especially when models can be mixed or hybrid in the sense that they represent not only causal dependencies, but both causal relations and supervenience relations. This suggests that the deeper problem leading to interventionist causal exclusion arises from the fact that variables standing in supervenience relations to each other violate what Woodward calls "independent fixability":

> 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)

If the set of variables V constituting a model satisfies independent fixability, then every combination of the values of the variables contained in V is metaphysically possible. This guarantees that it will always be metaphysically possible to change the value of one of the variables contained in V by an intervention and keep the values of all the other variables in V fixed. Does that mean that the interventionist exclusion problem can be blocked by requiring that the set of variables constituting a causal model must satisfy the independent fixability condition (for suggestions in this direction, see Polger, Shapiro, and Stern 2018; Yang 2013, 330)?[8]

The independent fixability condition implies that variables that stand in a supervenience relation to each other must not occur in the same causal model. Thus, if only variables included in the same model as X need to be held fixed by

[8] Note that in any case, independent fixability is only a necessary condition for the aptness of a model. Standard conditions often imposed on causal models are the causal Markov condition, the faithfulness condition, a minimality condition, and the requirement that the set of variables should be causally sufficient (Spirtes, Glymour, and Scheines 2000). Other possible conditions include the requirements that causal models should be monotonic (Hoffmann-Kolss 2024), and that the variables occurring in a causal model should have the right level of granularity and should have unambiguous effects on the other variables in the model (Woodward 2016). Since the primary goal of the present argument is to investigate what options the interventionist has to address the interventionist causal exclusion problem, it is sufficient for present purposes to discuss whether the variable set constituting a model should satisfy the independent fixability condition. Developing a complete theory of what conditions an apt model must satisfy would require considering aspects of causal modelling that are beyond the scope of the present argument.

an intervention on X, the requirement that the variables constituting a model must be independently fixable blocks interventionist exclusion. As pointed out in the previous section, however, the interventionist exclusion problem also arises with respect to non-mixed models, since the definition of intervention implies that not only variables occurring in the same model must be held fixed, but also that an intervention on X with respect to Y must be independent of all variables Z that are causally relevant to Y through a route that excludes X. Thus, simply imposing a ban on mixed models does not solve the interventionist exclusion problem.

In a paper also published in this volume, Weslake takes a slightly different route, arguing that the interventionist causal exclusion problem can be avoided by modifying the definition of intervention so that an intervention on X with respect to Y must be independent of all variables Z that are causally relevant to Y through a route that excludes X *in a model containing X, Y, and Z*. Models, in turn, must be constituted by sets of variables that satisfy the independent fixability condition. If X supervenes on Z, there is no model containing X, Y, and Z that satisfies the independent fixability condition, and interventions on X need not keep the value of Z fixed (Weslake 2024).

Weslake's proposed solution is technically superior to the exemption clause strategy discussed in the previous section. However, it comes at a cost. The independent fixability condition presupposes that the notion of possibility contained in it is interpreted in a certain way. Above, I interpreted it as metaphysical possibility, and this is also Weslake's reading (Weslake 2024). But this presupposes that there is a notion of metaphysical possibility and necessity that is at play in supervenience relations, and that it is strictly weaker than the notion of nomological necessity. The latter point is important. Otherwise, independent fixability would rule out too much. Consider, for example, the relation between the length of a pendulum and its period. Given that the period of the pendulum (T) is determined by its length (L) according to the formula $T = 2\pi\sqrt(L/g)$, one might conclude that L and T do not satisfy the independent fixability condition, since the values of L necessitate the values of T. But then imposing the independent fixability condition would imply that there can be no apt model that includes both variables L and T, and that the length of a pendulum would not qualify as a cause of its period. In general, if one could not distinguish between nomological necessity as the kind of necessity that applies to causal relations and a stronger notion of metaphysical necessity that applies to other metaphysical dependence relations, such as supervenience, grounding, or mereological relations, many causally related variables would violate the independent fixability condition (see Kistler 2013, 78–79), and thus would not be allowed to occur in the same causal model. But this is an unacceptable consequence.

At the end of the previous section, I argued that to solve the interventionist version of the causal exclusion problem, one has to employ metaphysically "thicker" notions or distinctions, rather than simply introducing exemption clauses for variables included in a supervenience base of the cause variable. The assumption that the condition of independent fixability excludes only metaphysically necessary dependence relations, whereas causal relations do not hold with metaphysical necessity (but only with nomological necessity), is an instance of this claim. Obviously, this approach is defensible, but it requires substantial and possibly controversial assumptions in the metaphysics of modality. In the next section, I will briefly discuss two alternative ways of dealing with multilevel causal structures.

6 Grounding and natural properties

An alternative possibility to address the problem raised by the exemption-clause strategy is to change the understanding of mixed models and stipulate that mixed models are ones that contain causal relations as well as grounding relations, where grounding relations are supposed to be metaphysically stronger than mere supervenience relations (for an application of this idea to the classical causal exclusion schema, see Kroedel and Schulz 2016; for arguments to the effect that the causal modelling framework can be applied to grounding relations as well, see Schaffer 2016; Wilson 2018). One can then argue that variables grounding the cause variable under consideration should be exempted from the requirement of having to be held fixed when intervening on the cause variable. Such an exception will plausibly not lead to the trivialization problem described in Section 4, because according to a standard theory of grounding, states or properties are not grounded in logically or mathematically more complex states or properties. For instance, the properties described by X will not be grounded in the properties described by X↔Y and X+Y.

Yet another approach to blocking the trivialization argument is to restrict the variables contained in an apt model to variables whose values satisfy a certain degree of naturalness, and to argue that variables of the form X↔Y or X+Y are not sufficiently natural. Like the previous two approaches, this approach relies on a possibly controversial metaphysical theory, in this case, a theory of properties, according to which properties or the states or events constituted by them can be ordered in terms of their degree of naturalness.

For reasons of space, I will leave open which of these two options is the most promising one. It should be noted, however, that there are contexts beyond the classical causal exclusion problem, where we have good reasons to work with mixed models, and to investigate what the adequacy conditions for such mixed models should be. For instance, causal relations occurring in the life sciences or the social sciences tend to be embedded in large causal networks considering a multiplicity of factors. Typically, these networks contain causes occurring at different levels. For instance, as soon as causal networks in the life sciences reach a certain level of complexity, they will usually contain causes and background conditions from lower levels, that is, the molecular or even the physical level. The fact that a model contains variables that describe entities belonging to different levels does not yet imply that the model is mixed. For instance, a very simple model containing only the variables M and P* (in the causal exclusion schema) would contain a cross-level *causal* relation, that is, the relation between M and P*, but would not be mixed, since it would not contain any variables standing in a supervenience or grounding relation to each other. However, many complex models covering different levels will contain supervenience or grounding relations in addition to causal relations.[9] This is a strong reason to extend classical causal modelling approaches to mixed models, and argues in favor of choosing a solution in terms of grounding or natural properties rather than the solution discussed in the previous section.

7 Conclusion

The upshot of this paper is that causal models that can account for multilevel structures cannot be had on the cheap. However the interventionist version of the causal exclusion problem is solved, the solution requires strong metaphysical assumptions about the aptness of causal models. It presupposes either strong assumptions about modality, that is, the distinction between metaphysical and nomological modality, or the notion of grounding, or the notion of naturalness applied to the values of the variables included in a causal model. The distinction between causes, confounders, and non-causally related variables is a metaphysically substantial one—and a causal modelling theory describing it has to be metaphysically substantial as well.

Acknowledgements

I would like to thank the editors of this volume, Katie Robertson and Alastair Wilson, and two anonymous referees for their very detailed and helpful comments on an earlier version of this manuscript.

This research was supported by the Swiss National Science Foundation (SNSF), grant no. 100019E_189589.

[9] A paradigm application is the mechanistic account of explanation in neuroscience (Craver 2007; Bechtel 2009).

References

Baumgartner, M. (2009), 'Interventionist Causal Exclusion and Non-Reductive Physicalism', *International Studies in the Philosophy of Science* 23(2): 161–178.

Baumgartner, M. (2010), 'Interventionism and Epiphenomenalism', *Canadian Journal of Philosophy* 40(3): 359–384.

Bechtel, W. (2009), 'Looking Down, Around, and Up: Mechanistic Explanation in Psychology', *Philosophical Psychology* 22(5): 543–564.

Blanchard, T. (2023), 'The Causal Efficacy of Composites: A Dilemma for Interventionism', *Philosophical Studies* 180(9): 2685–2706.

Bontly, T. D. (2002), 'The Supervenience Argument Generalizes', *Philosophical Studies* 109(1): 75–96.

Craver, C. (2007), *Explaining the Brain*, Oxford: Clarendon Press.

Eronen, M. I., and Brooks, D. S. (2014), 'Interventionism and Supervenience: A New Problem and Provisional Solution', *International Studies in the Philosophy of Science* 28(2): 185–202.

Gebharter, A. (2017), 'Causal Exclusion and Causal Bayes Nets', *Philosophy and Phenomenological Research* 95(2): 353–375.

Halpern, J. Y., and Pearl, J. (2005), 'Causes and Explanations: A Structural-Model Approach. Part I: Causes', *British Journal for the Philosophy of Science* 56(4): 843–887.

Hitchcock, C. (2001), 'The Intransitivity of Causation Revealed in Equations and Graphs', *Journal of Philosophy* 98(6): 273–299.

Hitchcock, C. (2007), 'Prevention, Preemption, and the Principle of Sufficient Reason', *Philosophical Review* 116(4): 495–532.

Hoffmann-Kolss, V. (2014), 'Interventionism and Higher-Level Causation', *International Studies in the Philosophy of Science* 28(1): 49–64.

Hoffmann-Kolss, V. (2022), 'Interventionism and Non-Causal Dependence Relations: New Work for a Theory of Supervenience', *Australasian Journal of Philosophy* 100(4): 679–694.

Hoffmann-Kolss, V. (2024), 'Bread Prices and Sea Levels: Interventionism, Monotonicity, and the Problem of Variable Relativity', *Philosophical Studies (special issue)*.

Kistler, M. (2013), 'The Interventionist Account of Causation and Non-Causal Association Laws', *Erkenntnis* 78(1): 1–20.

Kroedel, T. (2020), *Mental Causation. A Counterfactual Theory*, Cambridge: Cambridge University Press.

Kroedel, T., and Schulz, M. (2016), 'Grounding Mental Causation', *Synthese* 193(6): 1909–1923.

List, C., and Menzies, P. (2009), 'Nonreductive Physicalism and the Limits of the Exclusion Principle', *Journal of Philosophy* 106(9): 475–502.

McCain, K. (2015), 'Interventionism Defended', *Logos and Episteme* 6(1): 61–73.

Menzies, P. (2008), 'The Exclusion Problem, the Determination Relation, and Contrastive Causation', in J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*, Oxford: Oxford University Press: 196–217.

Polger, T. W., Shapiro, L. A., and Stern, R. (2018), 'In Defense of Interventionist Solutions to Exclusion', *Studies in History and Philosophy of Science Part A* 68: 51–57.

Raatikainen, P. (2010), 'Causation, Exclusion, and the Special Sciences', *Erkenntnis* 73: 349–363.

Rolffs, M. (2023), *Kausalität und mentale Verursachung: Eine Verteidigung des nicht-reduktiven Physikalismus*, Springer.

Schaffer, J. (2016), 'Grounding in the Image of Causation', *Philosophical Studies* 173(1): 49–100.

Shapiro, L. (2010), 'Lessons from Causal Exclusion', *Philosophy and Phenomenological Research* 81(3): 594–604.

Shapiro, L., and Sober, E. (2007), 'Epiphenomenalism—the Dos and the Don'ts', in G. Wolters and P. Machamer (eds.), *Thinking About Causes: From Greek Philosophy to Modern Physics*, Pittsburgh, Pennsylvania: University of Pittsburgh Press: 235–264.

Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction and Search*, 2nd ed, Cambridge, Massachusetts: The MIT Press.

Statham, G. (2018), 'Woodward and Variable Relativity', *Philosophical Studies* 175(4): 885–902.

Stern, R., and Eva, B. (2023), 'Antireductionist Interventionism', *British Journal for the Philosophy of Science* 74(1): 241–267.

Strevens, M. (2007), 'Review of Woodward, Making Things Happen', *Philosophy and Phenomenological Research* 74(1): 233–249.

Strevens, M. (2008), 'Comments on Woodward, Making Things Happen', *Philosophy and Phenomenological Research* 77(1): 171–192.

Weslake, B. (2024), 'Exclusion Excluded', in K. Robertson and A. Wilson (eds.), *Levels of Explanation*, Oxford: Oxford University Press: 101–135.

Wilson, A. (2018), 'Metaphysical Causation', *Nous* 52(4): 723–751.

Woodward, J. (2003), *Making Things Happen: A Theory of Causal Explanation*, New York: Oxford University Press.

Woodward, J. (2008a), 'Mental Causation and Neural Mechanisms', in J. Hohwy and J. Kallestrup (eds.), *Being Reduced: New Essays on Reduction, Explanation, and Causation*, Oxford: Oxford University Press: 218–262.

Woodward, J. (2008b), 'Response to Strevens', *Philosophy and Phenomenological Research* 77(1): 193–212.

Woodward, J. (2015), 'Interventionism and Causal Exclusion', *Philosophy and Phenomenological Research* 91(2): 303–347.

Woodward, J. (2016), 'The Problem of Variable Choice', *Synthese* 193(4): 1047–1072.

Woodward, J., and Hitchcock, C. (2003), 'Explanatory Generalizations, Part I: A Counterfactual Account', *Noûs* 37(1): 1–24.

Yang, E. (2013), 'Eliminativism, Interventionism and the Overdetermination Argument', *Philosophical Studies* 164(2): 321–340.

Zhong, L. (2020), 'Intervention, Fixation, and Supervenient Causation', *Journal of Philosophy* 117(6): 293–314.

6
From Multilevel Explanation to Downward Causation

David Yates

The principle that everything that happens within the physical domain has a sufficient cause within that same domain—the *causal closure of the physical*—poses a familiar causal exclusion problem for the special sciences: special science properties are distinct from their physical realizers, but if the physical domain is causally closed, then what causal work is left for such properties to do?[1] Here I argue that causal closure in fact poses no exclusion problem for the special sciences. I focus on the simple case of vector composition and argue that it involves irreducibly multilevel causation. Discussion of such simple physics may not seem like the most promising way of defending the autonomy of chemistry, biology, or psychology. My aim, however, is to persuade you that this case has profound implications for how we should think about causal closure. The way in which multiple vector fields compose, I shall argue, falsifies any closure principle according to which the course of physical events is entirely determined by properties at the fundamental level. I shall argue that the strongest closure principle that is consistent with vector composition allows for a particular form of downward causation, and so makes room for an irreducible causal role for special science properties. Hence, there is in principle plenty of causal work left for special science properties to do.

I shall assume an ontology of fundamental physical particles interacting in spacetime by means of fundamental physical forces. Macroscopic objects, I assume, are fully composed of such particles and macro-causal interactions are fully grounded in fundamental particle-particle interactions.[2] This might seem like a lot to assume, especially given that true fundamental physical reality may turn out to be very different. Why then do I set things up this way from the outset? One simple reason is that these are the terms of the causal exclusion debate. If particular special sciences entities are fully constituted by physical particles and their properties and relations, and everything that happens to those particles is fully determined by

[1] Kim (1992, 1998).

[2] For present purposes we can adopt a standard notion of grounding as a transitive, irreflexive, and asymmetric relation of metaphysical explanation holding between entities such as properties, states, objects, or events.

David Yates, *From Multilevel Explanation to Downward Causation* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © David Yates 2024. DOI: 10.1093/oso/9780192862945.003.0007

fundamental physical forces according to law, then what could special science entities be *doing*? And if special science entities have nothing to do, it seems just obvious that special sciences themselves are at most a simple and perspicuous way of grouping together fundamental physical causes. An ontology of fundamental particles, governed by a closed system of fundamental laws, is what gives rise to the exclusion problem in the first place. My aim here is to solve that problem on its merits: even assuming that everything is fully composed of physical particles, and that only fundamental physical forces are capable of accelerating such particles, there is still room for a kind of downward causation that would render special science autonomy unmysterious. Indeed, I shall argue, the idea that multiple fundamental forces compose to produce macroscopic effects *requires* downward causation of the kind in question.

What if it turns out—as many suppose it will—that the fundamental ontology of completed physics is radically different from that suggested by current theory? What if the fundamental ontology is not even spatiotemporal? Will the fundamental domain be in some sense causally closed, whatever causation turns out to be, and will that closure principle give rise to an exclusion problem? I confess that I don't know the answers to these questions, but since my intention here is to defend a solution to the exclusion problem as it is typically framed by those who think it is a serious problem worthy of attention, I don't see this as a major flaw. The reader may think of the proposed solution as follows: even if everything in our world were fully constituted by particles recognizably similar to those of current particle physics, and even if everything that happened to such particles were due to the exertion of fundamental forces similar to those we find in current physics, that would be no threat to the autonomy of the special sciences. In *that* kind of ontology—one, that is, in which the course of events is determined by the interaction of multiple vector fields in spacetime—downward causation is built in from the outset.[3] Now if it turns out that spacetime, particles, and forces are emergent and that, really, the fundamental dynamics have nothing to do with vector fields as we current conceive them, then the solution defended here is unlikely to solve exclusion problems that may arise between the fundamental ontology, whatever it turns out to be, and the emergent reality we seem to inhabit. But even if current physics is itself a special science, it's still interesting to consider whether the interlevel relations between physics and even higher-level special sciences bring with them problems of causal exclusion. If physics and everything above ends up being in some way excluded by whatever lies beneath, perhaps we can cross that bridge when we come to it.

Before proceeding, let me clarify the overall strategy of the paper. I intend to offer a solution to the exclusion problem that works by showing that causal closure does

[3] The central arguments of this paper go through *mutatis mutandis* in field-theoretic ontologies according to which fields are the fundamental constituents of physical reality, with particles derivative. I lack the space to defend that claim here and assume an ontology of fundamental particles for simplicity.

not rule out downward causation, so I will work with the strongest closure principle that could plausibly be justified by the available evidence. That closure principle alone, I shall argue, does not rule out downward causation and poses no problem for the autonomy of the special sciences. There is an even stronger closure principle in the vicinity, which I will also discuss, and which does pose such problems. However, I will argue, that principle is *falsified* by the available evidence rather than supported by it. The overall aim of my discussion is to explore the kind of downward causation that is consistent with the strongest plausible closure principle on offer, and thereby to cast light on what it is that special science properties could possibly be doing in order to earn their autonomy from physics. There are bound to be alternative perspectives on causal closure to the framework adopted here, and I do not claim to be solving any exclusion problems that may arise based on the resulting principles, however they may be defined. The proof, as always, is in the pudding: tell me what closure principle you are working with, why you believe it, and why you think it rules out downward causation as I shall conceive it here.

The plan of the paper is as follows. In Section 1, I first define and motivate the causal closure principle I shall be working with and then discuss the idea that causal closure entails causal-*explanatory* closure—the principle that everything that happens within the physical domain can in principle be fully *explained* without appealing to anything outside it. I do this because the case I consider involves irreducibly multilevel explanation and I want to draw conclusions from it concerning causation. I argue in Section 2 that causal explanations in physics that involve multiple forces are irreducibly multilevel, involving both force-generating basic physical properties and higher-level properties that partially determine how forces compose. In Section 3 I discuss the implications of multilevel explanations for the causal closure of the physical and suggest that the strongest empirically well-supported closure principle we can formulate is consistent with a certain kind of downward causation; I also discuss some options for understanding the metaphysics of this kind of downward causation in terms of causal powers. In Section 4 I discuss two potential sources of downward causation so conceived in terms of the debate between kinematic and dynamic theories of the origins of spacetime symmetries. I conclude in Section 5 with some reflections on special science autonomy.

1 From causal closure to causal-explanatory closure

Here is a typical way of stating the causal closure of the physical:

CC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t*.[4]

[4] In this formulation and those that follow, the notion of a sufficient cause may be read in probabilistic terms, i.e., as a cause that suffices to determine the chances of the effect.

According to CC, wherever in its causal history a given physical event has a cause at all, it has a sufficient physical cause. The content of 'physical' comes from physical theory, so that the closed domain is determined by our best fundamental science. There is undoubtedly more physics to be discovered and those yet-to-be-discovered things are holes in the causal structure of current physics. Nonetheless, I will assume that there is a complete theory in the vicinity and that it resembles current theory closely enough for CC to be non-vacuously true.[5] The intended scope of 'physical' in 'physical event' is not limited to the ontology of a completed physics. Rather, CC quantifies over the *broadly* physical, where an entity is broadly physical iff it is either part of the ontology of completed physics or appropriately related to such entities, where 'appropriate relations' include *grounding* relations such as composition, constitution, and realization.[6] The deflection of a particle in a magnetic field counts as a physical event under this definition and so does the eruption of a volcano.

CC is already quite a strong closure principle, but I will work with a much stronger one. Dialectically, the reason for this is as follows. My aim is to show that the strongest closure principle that is not falsified by evidence from simple physics still does not generate a causal exclusion problem for the special sciences. The reader may well suspect that the principle I eventually settle on is not actually supported by the available evidence, such is its strength. Fine with me—but if even that principle leaves room for downward causation, it should be clear that the weaker principles stated here—those which, the reader may suspect, have a chance of being true—do so as well. I shall thus formulate a sequence of closure principles of increasing strength, each one consisting in the preceding principle conjoined with an additional thesis. If a given principle in the sequence leaves room for downward causation, then so *a fortiori* do all those weaker than it. I note before proceeding that I do not claim that all possible closure principles can be ordered according to strength. As such, there may be principles that do not fit into the ordering developed here. The strongest true principle in my ordering does not give rise to an exclusion problem, but that entails nothing about whether exclusion problems arise from such hypothetical principles. No matter, for I am trying to solve the exclusion problem that arises from typical causal closure principles, not hypothetical problems that may arise from others.

[5] According to Hempel's dilemma, causal closure principles are either false (because indexed to current physics), or vacuous (because indexed to a future physics whose content is obscure). I am betting on future physics being sufficiently close to current physics to avoid the threat of vacuity. See Crane & Mellor (1990) for the dilemma; see Papineau & Spurrett (1999) for arguments that we don't need to define 'physical' to formulate contentful closure principles, provided we are clear about what the fundamental ontology will *not* include; and see Wilson (2006) for arguments that betting on future physics is the way to go, but with the addition of a 'no fundamental mentality clause' to rule out sui generis mental properties counting as physical.

[6] See Crook & Gillett (2001) for this way of thinking about the physical. Note that the scope of 'physical' in 'physical event' and 'sufficient physical cause' can be varied independently. I will do just that presently, in order to formulate the strongest closure principle I can.

An immediate worry with CC is that causal sufficiency is too metaphysically coarse-grained. Suppose events to be property-instances.[7] A physical event C might be causally sufficient for another such event E at *t* by means of an intermediary effect. Suppose C is synchronically sufficient for a *non*-physical event e*—an instance of a sui generis phenomenal property, say—and that it's only C and e* together that have the power to bring about E. This is an instance of traditional emergent downward causation, in which C synchronically causes an instance of a sui generis conscious property e* and e* then contributes a novel causal power, which (perhaps in combination with the powers of C) causally suffices at *t* for E. Given that CC does not rule this kind of situation out, we need a stronger closure principle. One way to secure such a principle is to add a clause stating that the physical cause in question has the power to bring about the effect in and of itself, in virtue of its physical properties alone. In the example sketched above, it's not the case that C's physical properties directly bestow upon C the power to cause E. Rather, they bestow upon C the power to cause e* and then the combined properties of C and e* bestow upon them the power to cause E. A causal power bestowal clause rules out this particular kind of emergent downward causation; I refer to the formulation below as *strong causal closure*:

> SCC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t and* C's physical properties bestow upon C all the powers needed to cause E.

According to SCC, every physical event that has a cause at any point in its history has a sufficient physical cause at that time, whose sufficiency is entirely in virtue of its physical features.[8] SCC is a *very* strong closure principle, but I propose to make it even stronger, by narrowing the scope of 'physical properties' to *basic physical properties*. What is a *basic* physical property? Recall that we are taking 'physical' to refer to entities that are part of the ontology of completed physics, or grounded by such entities. A basic physical property is a physical property whose instances are *not* grounded in their bearers' having any distinct physical properties. Examples may include electric charge, mass-energy, spin, entanglement, and spatiotemporal relations. In my usage, a *higher-level* property is any whose instances are grounded in the instantiation of further natural properties. This is a very weak condition on being a higher-level property, so weak that some such properties may well be within the ontology of completed physics: non-elementary values of quantitative

[7] For further details of the arguments that follow, see Yates (2009).

[8] I shall say more presently about the idea that a certain set of causal powers are *needed* to cause a given effect.

properties such as electric charge will come out higher level, for instance, because they are instantiated in virtue of other properties, in this case, having a certain number of particles of unit charge. Not all cases of higher-level properties are higher level in such an uninteresting way. Some higher-level properties, such as molecular geometry and neural firing patterns, are the proper subject matter of special sciences.

With this understanding of 'basic physical' properties in mind, this is the closure principle I shall be working with, which we may call *very strong causal closure*:

> VSCC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t* and C's basic physical properties bestow upon C all the powers needed to cause E.

Because C is a broadly physical event, the broadly physical particulars to which it happens are fully composed of fundamental physical particles. The idea behind VSCC is that the basic physical properties of these particles are responsible for C's power to cause E, in that they bestow upon C's basic physical proper parts all the powers that manifest as E's occurrence.[9] We can illustrate VSCC as follows. My raising my glass to take a drink of wine is fully grounded, on some occasion, by a certain plurality of basic physical events. Each of those events has a fully sufficient physical cause, in the sense that the power to cause each one is due entirely to the cause's basic physical properties. This then is the sense in which basic physical properties bestow all the powers that are *needed* for E to occur—given the combined manifestation of all those powers, *nothing else* needs to happen for me to sip my wine.

The central burden of this paper is to argue that *even VSCC* doesn't give rise to a problem of causal exclusion. Because my arguments depend on causal explanation, let's now turn to the issue of causal-*explanatory* closure. It seems obvious, reflecting on VSCC, that if we want to *fully explain* why some broadly physical event E happens, we do not have to appeal to anything outside the basic physical domain. Given that causal explanations cite causally relevant properties of causes, and all the causal powers C needs to bring about E are due to its basic physical properties, then it seems to follow that (whether or not we are able to formulate it) there is a complete causal explanation of E available in principle that cites only basic physical properties of its sufficient

[9] For indeterministic causation, we can think of sufficient physical causes as causes that suffice to fully determine the chances of occurrence of their effects. Note that VSCC as formulated here does not depend on the claim that particles are basic physical and is entirely consistent with a field-theoretic ontology in which particles are metaphysically grounded. I lack the space here for a detailed treatment, but the central arguments of this paper will go through *mutatis mutandis* if (as seems to be the case) fields are more fundamental than particles.

cause C. Thus, it seems that VSCC entails the following principle of causal-explanatory closure:

> CEC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t* and E can be fully causally explained in terms of C's basic physical properties.[10]

Let's say that a *full* causal explanation is one that cites all the causal work that was done to bring about the effect—whatever contributed to an effect's occurrence, a full causal explanation thereof will refer to it. It's now fairly easy to see why CEC poses a problem for the autonomy of the special sciences. On the assumption of physicalism, special sciences such as chemistry, biology, and psychology are in the business of explaining broadly physical events and processes—such as bonding, digestion, and cognition. But according to CEC, all broadly physical events can in principle be fully explained without leaving the basic physical domain. Given CEC, it seems that if special sciences are autonomous, it is not in virtue of what they explain, but the manner in which they explain it. A popular strategy for defending special science autonomy is to accept this consequence of CEC and attempt to give an account of why special science explanations of certain phenomena are better than the corresponding basic physical explanations, which are assumed to be in principle available.[11]

Despite the appearances, CEC does not follow from VSCC. In the brief argument I gave above for CEC based on VSCC, I assumed that all causally relevant properties of C in relation to E—that is, all the properties a full causal explanation of E in terms of C must cite—are properties that bestow upon C the power to cause E. In other words, I assumed that the causal work that properties do in relation to some effect *consists in* bestowing the power to cause it. Given this assumption, VSCC entails that basic physical properties *do all the causal work* involved in C's causing E, which in turn implies that we can fully explain E in terms of C. We can state the assumption required to derive CEC from VSCC as follows:

> CW: all the causal work that properties do consists in causal power bestowal.

[10] Things are somewhat complicated here if E's cause C only brings about E by sufficing to determine its probability. In cases where multiple outcomes are equally probable, the cause won't explain why the actual outcome occurred rather than some other. However, where there is truly indeterministic causation, explanations of this kind are the best we can do and full in the sense that there is nothing missing from them.

[11] Versions of this strategy can be found in Fodor (1974, 1997); LePore & Loewer (1987); Jackson & Pettit (1990); Yablo (1992); List & Menzies (2010); Wilson (2011); and Yates (2012). Kim (1998) regards all such accounts as a free lunch and I am now inclined to agree, which is why I am here defending downward causation.

VSCC says that C's basic physical properties bestow all the causal powers required to cause E. If that's all the causal work there is to do in relation to E, it follows that there is a full causal explanation of E in purely basic physical terms. The conjunction of VSCC and CW entails CEC. We can think of this conjunction as defining a causal closure principle, which we may call *super-strong causal closure*:

> SSCC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t* and C's basic physical properties bestow upon C all the powers needed to cause E; and all the causal work that properties do consists in causal power bestowal.

SSCC entails CEC and for that reason is just the kind of closure principle needed to undermine the autonomy of the special sciences. It is also plausibly the kind of principle Kim has in mind when he insists that closure entails that there is *no causal work left* for special science properties to do. However, as I shall now argue, CEC is falsified by some very simple physics, hence SSCC is false. It follows in addition that either (1) VSCC is false, or (2) CW is false. If we choose (1), then there are causal powers beyond those that basic physical properties bestow; if we choose (2), there is more to causation than causal power bestowal. Either way, I shall suggest, causal closure poses no problem for the autonomy of the special sciences. In the remainder of this paper, I will assume that VSCC is true and develop a position based on option (2), the idea that there is more genuine causal work to be done even after all the relevant causal powers have been bestowed.[12]

2 The role of geometric structure in vector composition

When philosophers rely on scientific cases to justify claims of emergence or downward causation, they typically pick sophisticated cases like ferromagnetic phase transitions in condensed matter physics, or self-organization in systems biology. By contrast, the case considered here is simple, but I think it tells us a lot about causation, causal powers, and causal closure. In fact, I think it gives us very good reason to believe in a kind of downward causation, which in turn calls for a metaphysical explanation. The benefit of simplicity is that if I'm wrong about all this, it should be easy for readers to see why.

In this section, I will try to convince you that downward causation is as ubiquitous as vector composition. I shall leave the task of explaining how this kind of downward causation is possible to Sections 3 and 4, where I will discuss its metaphysical nature and ultimate source, respectively, in order to render it

[12] I defended a version of (1) in relation to molecular geometry in Yates (2016), but now prefer (2), for the reasons given in Section 3.

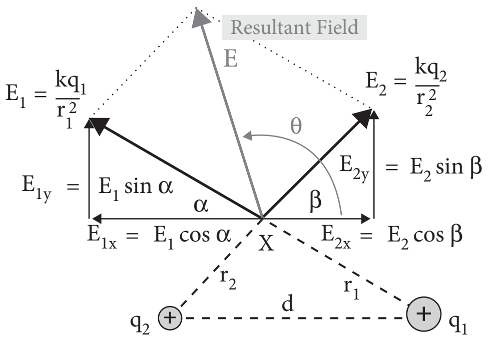
Resultant Field
E
$E_1 = \frac{kq_1}{r_1^2}$
$E_2 = \frac{kq_2}{r_2^2}$
θ
E_{1y} = $E_1 \sin \alpha$
E_{2y} = $E_2 \sin \beta$
α
β
E_{1x} = $E_1 \cos \alpha$
X
E_{2x} = $E_2 \cos \beta$
r_2
r_1
d
q_2
+
+
q_1

Figure 6.1 Calculation of the electric field E due to two point charges.

unmysterious and unproblematic. The result, I hope, will be a position that shows how something close to traditional emergent downward causation—close enough to defend the autonomy of the special sciences—is possible within a robustly reductionist metaphysical framework.

Figure 6.1[13] shows the calculation of the resultant field E due to two point charges, q_1 and q_1, at a distance r_1 from q_1 and r_2 from q_2, where the dotted lines from the two charges meet, marked 'X'. This is an illustration of the *parallelogram rule.*

The calculation proceeds by resolving E_1 and E_2 into their horizontal and vertical components. The x and y components of the resultant field E are given by:

- $E_x = E_{1x} + E_{2x}$
- $E_y = E_{1y} + E_{2y}$

The magnitude and direction of the resultant field E are then given, respectively, by:

1. $E = \sqrt{\left(E_x^{\,2} + E_y^{\,2}\right)}$
2. $\tan\theta = \left(E_y / E_x\right)$

Suppose this calculation forms part of the explanation of the acceleration of a charged particle located at X. My central claims here are that (1) the explanation in question is *irreducibly multilevel*, which violates CEC; and that (2) this

[13] Reproduced with permission from url: http://hyperphysics.phy-astr.gsu.edu/hbase/electric/mulpoi.html#c3.

violation is due to a kind of downward causation, whose consistency with VSCC depends on how we think about causal powers. More on that presently. Figure 6.1 depicts both basic physical and higher-level properties of the two charges. It shows (A) the specific spatial relations that obtain between q_1 and q_2 and the point X at which we want to calculate the resultant (being separated by r_1 and r_2, respectively); and the distance d between the particles. But in addition to this, it shows (B) what I call the *geometric structure* of the particles in relation to X, in this case given by the angles α and β that we used to factor the component vectors into their horizontal and vertical components. Why isn't geometric structure basic physical? Simply put, because the geometric properties in (B) are instantiated *in virtue of* the basic physical properties given in (A) and hence grounded therein. The specific spatial relations in (A) are not the only way to achieve the geometric structure in (B). We can vary r_1 and r_2 independently while holding α and β fixed, resulting in the same geometric structure. Thus, geometric structure is multiply realizable.[14] I shall now argue that in the present case, geometric structure has a distinctive and irreducible causal-explanatory role compared to the basic physical properties that realize it.[15]

As noted above, we can vary basic physical properties without changing geometric structure. Suppose we move q_1 away from X without changing α. The x and y components of E_1 will decrease in magnitude as $1/r_1^2$, but they will remain in constant proportion to each other, since the direction of E_1 will remain the same. This change will alter both the direction and magnitude of the resultant field E, in a manner given by equations (1) and (2) above. This shows that the basic physical spatial relations of the particles make a difference to the resultant field independently of the geometric structure they realize. However, *the converse is also true*. Imagine now moving q_1 towards q_2 in a circle of radius r_1 about X. This won't change the magnitude of E_1, but as α varies, its direction will change, resulting in different values for E_{1x} and E_{1y}, hence altering the magnitude and direction of E according to equations (1) and (2). This, then, is a way for geometric structure to make a difference independently of the values of r_1 and r_2, which remain constant. There is a complication, however. Because geometric structure is realized by, hence supervenient on, basic physical properties, it isn't

[14] This kind of multiple realizability—in which a property P is instantiated in virtue of basic physical properties and there are many different configurations of basic physical properties that are sufficient but not necessary for P—won't count as genuine multiple realization for those, like Shapiro (2000), who see multiple realization in terms of structurally heterogeneous ways of implementing a given function. For others, such as Gillett (2003), multiple realization is easier to come by. I am with Gillett on this, but won't take a stand on the issue here, as my arguments only require that geometric structure is a higher-level property.

[15] What follows is a simplified and (I hope) improved version of an argument I gave in Yates (2016), to show that the molecular geometry of water plays an irreducible role in determining its dipole moment.

possible to change the former without some change in the latter. It might be suspected, then, that it is this basic physical change that is the real difference-maker in respect of E.

It's important to get clear about which basic physical changes occur. We start out with q_1 and q_2 located a distance r_1 and r_2 from X respectively and a distance d apart. As we move q_1 towards q_2, it remains r_1 from X and q_2 remains r_2 from X. The only basic physical parameter we change is the distance between q_1 and q_2. But there's nothing special about moving the particles from d to d' apart that explains *why* the magnitude or direction of E changes in the way that it does. Rather, what explains the changes in E is that we move q_1 towards q_2 *holding fixed r_1 and r_2*. In order to do that, we have to change the geometric structure of q_1, q_2, and X. It's only given r_1 and r_2 that a certain separation between q_1 and q_2 determines the magnitude and direction of E, because it's only *together* that these three parameters suffice to determine a geometric structure. The basic physical property we have to change in order to change the geometric structure makes a difference to the resultant field only *because* it makes a difference to the geometric structure. In and of itself, the basic physical change alone doesn't explain the change in the resultant.

What follows from this? I think the example shows that dynamical causal explanations that feature vector composition are irreducibly multilevel and hence that CEC is false, from which it follows that SSCC is also false. An explanation of the acceleration of a charged particle at point X must involve not only basic physical properties such as the charges of the particles q_1 and q_2 and the spatial relations r_1 and r_2, but also an ineliminable appeal to the geometric structure formed by q_1, q_2, and X. There's a simple reason for this: part of the explanation of the magnitude and direction of E is the degree to which the fields due to q_1 and q_2 point in the same direction at X. And *pointing in the same direction* is an irreducibly geometric property. If I am correct that geometric structure is grounded in basic physics and multiply realizable, then an entire family of familiar, simple dynamic explanations involve both basic physical and higher-level properties. Crucially, the higher-level properties are not merely a proxy for their basic physical realizers. As we've seen, in the case of the difference-making role of geometric structure, the basic physical properties that realize that structure are explanatory only insofar as they realize the geometric structure in question. This doesn't settle the issue of what it is that geometric structure actually does, or how it does this—thus far I have only argued that CEC and hence SSCC are false, leaving open the following options: (1) geometric structure is itself a powerful property, which would violate VSCC as well; or (2) geometric structure does novel causal work without bestowing novel powers and hence without violating VSCC. In the next section I defend option (2) and say more about the kind of downward causation involved.

3 What kind of downward causation?

To facilitate the discussion that follows, it will be useful to have the relevant principles of Section 1 to hand. Recall that I am working with the following causal closure principle:

> VSCC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t* and C's basic physical properties bestow upon C all the powers needed to cause E.

The conjunction of VSCC with the following principle about causal work,

> CW: all the causal work that properties do consists in causal power bestowal,

entails a principle of causal-explanatory closure:

> CEC: every physical event E that has a cause at *t* has a sufficient physical cause C at *t* and E can be fully causally explained in terms of C's basic physical properties.

If the arguments of Section 2 are correct, then CEC is false, so we must reject either CW or VSCC. I shall discuss both strategies in what follows. If we say that the causal-explanatory role of geometric structure stems from its bestowing novel causal powers, then we must reject VSCC. This strategy seems plausible if we have an antecedent commitment to CW. If a property's causal-explanatory role stems solely from its bestowing causal powers on its bearers, then if a property P is one without which a full causal explanation of some token effect can't be given, P must bestow a novel causal power to bring about that effect on that occasion. One way to make sense of properties like geometric structure bestowing powers is to think in terms of Shoemaker's *conditional* powers.[16] For x to have the power simpliciter to ϕ is for x to be disposed to ϕ, under certain conditions C. For x to have a conditional power to ϕ is for x to be such that if it had certain other properties, it would have the power to ϕ simpliciter, where the other properties in question are not independently sufficient for this.

As Shoemaker notes, conditional powers enable us to isolate the causal contributions of individual properties to a power simpliciter when that power is jointly bestowed by several properties. It's plausible that basic physical properties such as electric charge bestow certain powers simpliciter, but that doesn't seem to be the case with geometric structure. There is no obvious power simpliciter that

[16] Shoemaker (2001), pp. 25–26.

all n-tuples with the same geometric structure possess. However, the idea that spatiotemporal properties like geometric structure bestow *conditional* powers is more compelling. The difference-making role of geometric structure described in Section 2 suggests that it determines the extent to which the fields of two charged particles are cooriented at X, which in turn determines a range of possible resultant fields, of different magnitudes and directions. We might then say that geometric structure bestows upon q_1 and q_2 the power to accelerate a positively charged particle located at X at a certain rate in the direction of the resultant E, conditionally on the values of q_1, q_2, r_1, and r_2. Note that a corresponding conditional power is also bestowed on q_1 and q_2 by their charge, in this case conditionally on r_1, r_2, and geometric structure. On this approach, geometric structure bestows novel conditional powers, in line with the novel difference it makes to the resultant field. This kind of novelty resembles strong emergence, traditionally conceived, for we have a dependent property with powers that are not inherited from its physical realizer.[17]

The alternative is to deny CW and keep hold of VSCC. On this approach, all the causal *powers* involved in causing a physical effect are bestowed by the basic physical properties of a sufficient physical cause, but other properties may also be involved in *causing* the effect in question. The causal powers that manifest when a particle accelerates in the resultant field of multiple charged particles are bestowed by electric charge, but geometric structure is involved in determining the direction of the resultant field and hence in determining *how* the powers in question manifest on some occasion. We are now faced with the question of how this kind of causal work relates to powers and their manifestations. For present purposes, the following simple account will suffice. Why not simply say that geometric structure is among the *manifestation conditions* of the powers (simpliciter or otherwise) of the two charged particles? On this account, all causal powers are due to basic physical properties, but at least some such powers have irreducibly geometric conditions on their manifestation.

It is widely held that causal powers have manifestation conditions—there are certain things that need to happen for a power to produce its characteristic effect. The power of a knife to cut butter, for instance, will manifest if the knife and butter are brought into contact in the right way. A simple way of thinking about the arguments of the present paper is then as follows: some powers are such that what you have to do to get them to manifest is arrange their bearers in a certain kind of geometric pattern. On this interpretation, we may say that q_1 and q_2 have the power to accelerate a positively charged particle located at X at a certain rate in the direction of the resultant E, that they have this power entirely in virtue of their basic physical properties, but that if you want to get the power

[17] This was the strategy I employed in Yates (2016) to argue that the causal novelty traditionally attributed to strongly emergent properties is consistent with physical realization.

in question to *manifest*, you have to fix r_1, r_2, and the relevant geometric structure. Given that vectors compose parallelogram-wise, the idea that geometric properties might be among the manifestation conditions of fundamental forces is, I think, rather compelling.[18]

We have seen two ways of making sense of the causal-explanatory novelty of geometric structure in vector composition. We can either: (1) say that geometric structure bestows novel conditional powers in line with its novel difference-making role and because of this violates VSCC; or (2) all causal powers are due to basic physical properties, but geometric structure is among the manifestation conditions of such powers, which in turn violates CW but leaves VSCC intact. There is at least one reason to prefer option (2). If we choose (1), there is an odd symmetry between what electric charge does and what geometric structure does, since both are regarded as bestowing conditional powers, which combine to yield a power simpliciter. But that is to gloss over a significant difference in the contributions of these properties—electric charge also bestows powers simpliciter, whereas geometric structure does not. There is no power that all bearers of a certain geometric structure have in common, regardless of their other properties. Since we are clearly already committed to causal powers due to properties such as electric charge, it would be extravagant to posit geometric conditional powers as well if we can understand the causal role of geometric structure in terms of just the former. In what follows, for brevity I will refer to the causal role of being a manifestation condition as *conditioning*.

It might be objected that conditioning isn't really a causal role, so downward conditioning—a higher-level property being a condition on the conditional powers of a basic physical property—isn't really downward *causation*. The conditioning role in question doesn't violate the causal closure of the basic physical domain as formulated in VSCC, because it doesn't consist in power bestowal. But if some of the powers of basic physical properties have irreducibly higher-level manifestation conditions, then there is more causal work to do even after all the causal powers have been bestowed. If you don't want to call it causal work, then call it something else. The fact remains, higher-level properties like geometric structure have a conditioning role that they don't inherit from their realizers, which role is among the determinants of the dynamic evolution of basic physical systems. That looks like downward causation to me. I turn now to the question of its source.

[18] There is another option that I do not consider here for reasons of space, which is to take geometric properties as extrinsic conditions under which some of the conditional powers bestowed by basic physical properties become powers simpliciter. It is somewhat controversial to hold that there is a principled distinction between ordinary manifestation conditions and extrinsic conditions on conditional powers. I defend that claim in order to develop a geometric version of hylomorphism in Yates (forthcoming).

4 Constraint or coincidence?

The source of downward causation, as understood here, depends on what we say about the source of the law of vector composition. Lange distinguishes two potential sources for explaining why forces compose parallelogram-wise.[19] There are both dynamic and static explanations available, which is to say that the parallelogram law can be deduced either by considerations that stem from the dynamical laws (by considering how the displacements that forces produce compose) or from statics (by considering how general symmetry principles impose vector addition on the force laws). In the first case, vector composition of forces is grounded in the fact that forces produce accelerations and accelerations compose parallelogram-wise. In this case, there is no unifying principle that explains why other vectors—electric fields, gravitational fields, temperature gradients, heat flows, etc.—also compose according to the parallelogram rule. In the second case, the law of vector composition is a constraint on the dynamics and has its source in symmetry principles that are independent of the precise forms of the dynamical laws. These symmetries can also be applied to the case of other vectors, *mutatis mutandis*, so there is the possibility of a unifying explanation.

Lange doesn't take sides in this debate, since his primary aim is to argue that in order to make sense of the debate itself, we need to appeal to a nested hierarchy of laws. Roughly, he suggests that a law L is a constraint iff L would still have held even if the dynamical laws had been different. Consider the counterfactual 'had the dynamical laws been different, vector fields would still have composed parallelogram-wise'. If the law of vector composition holds in virtue of the form that the dynamical laws happen to take at our world, then this counterfactual comes out false—we have no right to assume that a coincidence of the dynamics would still have held on the counterfactual supposition that the actual dynamical laws don't hold. Conversely, Lange argues, for this counterfactual to be true is for the relevant symmetry principles to be *more necessary* than the dynamical laws—to hold, that is, however the dynamical laws may be.

Lange considers dynamic vs static explanations of the parallelogram rule as part of an argument that a particular historical debate in foundations of physics is best understood in terms of a nested hierarchy of laws. However, it should be noted that the main static account he considers relies on symmetry principles that might themselves be explained dynamically. In Poisson's account of the parallelogram rule, the main symmetry principle involved is spatial isotropy—the principle that space has no preferred direction. Poisson's explanation is complex, and I lack the space to give a full account, but the first stage of the proof relies on isotropy to establish the magnitude and direction of the resultant force of two equal forces.[20]

[19] Lange (2017), Ch. 4.

[20] Lange (2017), pp. 167–169.

He then generalizes the proof to unequal forces. But as Lange himself notes elsewhere, the same question—constraint or coincidence—can be raised in relation to symmetry principles such as isotropy.[21] Brown and Pooley argue[22] that the symmetries of Minkowski spacetime, including isotropy, are grounded in the dynamics. As they see it, Minkowski spacetime is just our best codification of the primitive Lorentz covariance of the dynamical laws: for Minkowski spacetime to have the metric properties and symmetries that it does is for the dynamical laws to be Lorenz covariant.[23] There's no unifying explanation, in Brown's view, as to why all the laws are Lorentz covariant, so the Minkowski metric arises from a coincidence of the dynamics.[24] By contrast, a constraint-based explanation of spacetime symmetries will need to locate their source somewhere other than in the dynamical laws—for instance in the structure of spacetime itself. I shall now focus on the implications of this debate for the source of downward causation.

Let's begin with the dynamical account, according to which the dynamical laws are primitively Lorentz covariant, which determines that the Minkowski metric is the best way of codifying those laws. Spacetime symmetries are fundamentally symmetries of the dynamics, so vector composition is likewise a consequence of the dynamics. It is, on this view, a coincidence that the various vector quantities all compose parallelogram-wise. The dynamical account leads to what might be termed 'upward-downward causation'. The symmetries of the fundamental force-laws determine that multiple forces compose according to the parallelogram rule, which is to say the laws determine that geometric structure plays a novel role in determining the resultant of multiple forces. Downward causation, on this view, is written into the fundamental dynamical laws. And if one thought, as dispositional essentialists do, that the dynamical laws themselves are grounded in the essences of basic physical properties, then those very essences would explain why certain non-basic properties like geometric structure play a novel role in determining the course of events. We can think of it in terms of the relational individuation of powers. In a pure powers ontology, it is typically held that basic physical properties are individuated solely by type-level causal relations they bear to each other. What the case of vector composition shows is that such properties are individuated by

[21] Ch. 3 of Lange (2017) is devoted to this issue.

[22] Brown & Pooley (2001); Brown (2005); Brown & Pooley (2006).

[23] Lorentz covariance of laws is the property of being invariant under the Lorentz transformations of special relativity, which tell us how to transform the coordinates of point-like events between inertial frames. The laws of physics are Lorentz covariant, which is to say they are the same in all inertial frames, subject to Lorentz transformation of the relevant coordinates. See Lange (2017), Ch. 3, for a detailed account.

[24] Brown (2005), p. 143; see also Lange (2017), pp. 112–113. Brown's position is anti-substantivalist about spacetime and for this reason is typically seen as a form of relationism. See Knox (2019) for full discussion and for a functionalist interpretation of the dynamical conception of spacetime as applied to general relativity.

relations they bear both to each other *and* to higher-level geometric conditions on their composition.[25]

As noted above, on the dynamical explanation it's not true that had the dynamics been different, vectors would still have composed parallelogram-wise. If the parallelogram rule is a coincidence, then we have no right to suppose that the dynamical laws at all the relevant worlds have the same symmetry properties as the actual laws, for the simple reason that coincidences of the actual dynamics are not robust under the counterfactual supposition that the actual dynamical laws don't hold. Conversely, for Lange, if the parallelogram rule still holds under the counterfactual supposition that the dynamical laws don't, then it must have a higher grade of necessity. Lange appeals to laws that hold at a broader range of possible worlds than the dynamical laws, which in Lange's system are understood in terms of a nested hierarchy of primitive subjunctive facts. In the remainder of this section, I shall offer an alternative to Lange's view, which treats vector composition as a constraint within an essentialist framework and hence has no need for primitive subjunctive facts.

The Lorentz transformations can be derived from purely kinematic principles, independently of the dynamics.[26] Lange offers such a derivation, from the principle of relativity (normally stated as the claim that the laws of physics take the same form in all inertial frames) and the invariance of the spacetime interval.[27] What's important for present purposes is that spacetime symmetries such as isotropy are here assumed by way of explaining why the dynamical laws are Lorentz covariant. As Lange notes, the principle of relativity entails spatiotemporal isotropy and homogeneity, so the kinematic explanation builds in a lot of spacetime structure at the outset. One way of interpreting this is to say that the relevant spacetime structure belongs to a substantival Minkowski spacetime and that the nature of spacetime itself is what grounds Lange's constraints. Lange himself does not commit to substantivalism, preferring instead to treat the laws as primitive subjunctive facts; for this reason, his own position is consistent with a thin conception of spacetime similar to Brown's but one which treats its structure as grounded in primitive *kinematic* laws rather than primitive dynamical laws.[28] To say that

[25] I develop this idea in detail in Yates (2018) and argue that it helps pure powers ontologies avoid a regress.

[26] Janssen (2009); Lange (2017), section 3.2.

[27] Lange uses a kinematic version of the relativity principle, stated as follows: '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' (2017, p. 104). It's more common for derivations of the Lorentz transformations to appeal to the principle of relativity together with the light postulate—the principle that, as measured in any inertial frame, the speed of light is a constant regardless of the motion of the source. The details of these derivations need not concern us here.

[28] Heron & Knox (2019) make this point. Like Lange, Janssen (2009) also eschews any commitment to spacetime substantivalism, despite claiming that spacetime structure explains the Lorentz covariance of the dynamical laws. In Janssen's case things are more complex, as he does not have Lange's nested hierarchy of laws available to render relationism consistent with his view. See Acuña (2016) for arguments that Janssen is tacitly committed to substantivalism about Minkowski spacetime.

spacetime symmetries are constraints is then to say, for Lange, that the kinematic laws have a higher grade of necessity than the dynamical laws.

For those who prefer to ground laws in essences, substantivalism offers an attractive alternative. If symmetry properties like isotropy are among the essential properties of a substantival spacetime, then not only will the nature of spacetime ground the parallelogram rule, but we can also make sense of constraints without positing degrees of necessity or primitive subjunctive facts. In the arguments that follow, I assume no particular version of spacetime substantivalism and do not commit to the view that spacetime points are primitive objects. In other words, I assume that there are versions of substantivalism available that can avoid the hole argument.[29] The only aspect of substantivalism that is required in the present context is the claim that spacetime is ontologically independent of spacetime occupants. This minimal substantivalist claim is consistent with all forms of substantivalism, including metrical essentialism, according to which spacetime points are individuated by their metric properties; and spacetime structuralism, if indeed this is importantly different from metrical essentialism. We need not embrace problematic haecceitistic spacetime points in order to embrace substantival spacetime, provided we do not deny the reality of spacetime points altogether.[30]

Consider again the counterfactual: 'had the dynamical laws been different, vector fields would still have composed parallelogram-wise'. Given substantivalism, it now makes sense to suppose that the nearest worlds at which the dynamical laws are different are worlds at which spacetime structure is, insofar as possible, the same. For instance, if we are focused on worlds at which Coulomb's law is an inverse cube law, then we should think of worlds with one extra spatial dimension, but not worlds at which spacetime is not isotropic. Holding as much as possible of our spacetime structure fixed, we can imagine varying the dynamical laws in certain ways, since spacetime itself—as far as we know—does not suffice to fix those laws. On a dispositional essentialist account of the dynamical laws, the ontological independence of spacetime gives rise to an interesting asymmetry. If the dispositional essences of basic physical properties such as electric charge are given in terms of Lorentz covariant dynamic equations, then those properties will be ontologically dependent on spacetime. If spacetime structure doesn't depend on what occupies it, then the dependence is asymmetric.

[29] See Norton (2019) for an introduction to the hole problem. The problem arises due to diffeomorphism invariance in general relativity (GTR). We can assign different metric properties to spacetime points within a particular region (the 'hole') leaving points outside the same, preserving all observational consequences of GTR. But bodies within the hole move along different trajectories, so it looks like GTR violates determinism, because what goes on outside the hole doesn't fix what happens inside it. Whether or not hole diffeomorphisms represent genuinely different physical possibilities depends on how we conceive of spacetime points.

[30] Maudlin (1989); Bartels (1996). This is also a feature of Pooley's (2006) sophisticated substantivalism.

Things are somewhat different in certain approaches to quantum gravity—in loop quantum gravity, for example, spatiotemporal localization is understood in terms of the interaction of fundamental quantum fields with a quantum gravitational field. On this view, the fundamental matter fields, the electromagnetic field, and the gravitational field might all be mutually ontologically independent, with dynamical laws arising as a result of the interaction between them.[31] Even assuming an eventual quantum theory of gravitation, what's important for my purposes is that provided spacetime is not ontologically dependent on the other quantum fields, its structure will be counterfactually robust enough to ground constraints without primitive subjunctive facts. It's not that the parallelogram rule is *more necessary* than the dynamical laws, on this approach—it's indexed to a different portion of modal reality because it has a different *source*. This in turn has the consequence that the closest possible worlds at which the dynamical laws don't hold are worlds at which a substantival spacetime with the same symmetry properties as ours exists, hence worlds at which constraints like the parallelogram rule, which follow from spacetime symmetries, also hold.

If vector composition has its source in spacetime structure, then downward causation flows from the way in which spacetime itself constrains the dynamics. However, we must be careful not to conflate the present sense of 'constraint' with a causal notion according to which spacetime literally forces bodies to follow inertial paths and forces vector fields to compose parallelogram-wise. Brown objects to this suggestion on the grounds that it imports causal powers to spacetime and 'spacetime feelers' to material bodies.[32] If spacetime constrains the dynamics causally, then it has the power to cause a body moving in the absence of forces to follow an inertial path; and it seems the body itself must then have the reciprocal power to be so constrained—the power to 'feel' which way the inertial paths point. Spatiotemporal constraints, however, were supposed to be independent of the dynamics of spacetime occupants. Whatever it is for spacetime to constrain the motion of a body or the propagation of a field, it seems it cannot be understood as the manifestation of a power. How then should it be understood?

I do not here claim that spacetime is a cause; rather, the claim is that geometric properties have their causal roles in virtue of the essential nature of spacetime. It might seem as though there is a conflict here between a non-causal notion of constraint and a causal role for geometric structure, but it's one thing to talk about

[31] For more on the potential philosophical implications of a quantum theory of gravity, see the essays in Wüthrich, Le Bihan, & Huggett (eds.) (2021). It is hard to know what will become of properties such as mass and electric charge, not to mention the idea that they bestow causal powers, in quantum gravity. Some such theories seem to have the consequence that spacetime itself is in some sense emergent rather than basic physical, and if so then presumably causal powers, understood as, e.g., powers to produce accelerations, will also be emergent. In principle, however, I think the solution to the exclusion problem presented here could be reinterpreted as way of explaining why one special science—current physics—does not causally exclude those above.

[32] Brown (2005).

spacetime itself having causal powers and another thing entirely to talk about the causal role of geometric structure conceived as a property of pluralities of spacetime *occupants*. The kinematic explanation of why the parallelogram rule holds is not causal, but given that the rule does hold, instances of geometric structural properties play an irreducible causal role in determining the dynamic evolution of basic physical systems. This is a form of downward causation not because spacetime itself exerts a causal influence on things that are beneath it in a levels hierarchy—indeed, if spacetime itself is a basic physical structure, then there won't be anything beneath it—but because instances of higher-level geometric properties exert such an influence by acting as manifestation conditions on the causal powers of basic physical properties. Spacetime substantivalism is a potential explanation of the *source* of geometric downward causation, but that, as noted above, should not be taken to imply that spacetime itself is a *cause*.[33]

5 Conclusion: so what?

The kind of downward causation defended here may not seem sufficient to defend the autonomy of the special sciences, so let me say something in conclusion about why I think it's a promising start. Firstly, whether it's enough for a robust positive defence, it is sufficient to undermine arguments against special science autonomy based on closure principles such as VSCC. The case of vector composition leaves us with two options in relation to closure: either (i) the basic physical domain is not causally closed because geometric structure bestows novel conditional powers; or (ii) the basic physical domain is causally closed, but geometric structure still has a novel role to play as a condition on the manifestations of basic physical powers. If the causal closure of the basic physical domain is false, then it doesn't pose a problem for the special sciences; if it's true but consistent with an irreducible causal role for higher-level properties, then it is once again consistent with special science autonomy. This does not suffice to show that the special sciences *are* autonomous from physics, but it does serve to undermine the strongest argument that they are not.

Secondly, the arguments presented here can also be used to defend the claim that the special sciences are indeed autonomous. One of the core things that special sciences such as chemistry and biology do is to classify by spatiotemporal properties such as geometric structure, so their autonomy is secured by dint of the downward causal influence of such properties on the dynamics of basic physical systems. Chemistry, for instance, classifies molecules in terms of properties such as being linear, planar, bent, cyclic, and so forth, all of which are geometric. If I am right

[33] I thank Katie Robertson for pressing me on this issue.

that properties such as these are causally novel, then there is no threat to chemistry from below. Chemical properties such as molecular geometry are just the right kind of properties to condition the powers bestowed by their basic physical realizers. This simple account may not extend in a natural way to other special sciences, but I see no reason in principle why conditioning by higher-level properties should not occur in sciences such as neuroscience[34] and psychology as well, even if it is not conditioning of the kind that we see in vector composition. The devil is no doubt in the details, but if at least some special science properties have a genuinely novel downward causal role in relation to physics, then that is a promising start.

Traditional accounts of special science autonomy are often framed within a functionalist conception of special science properties. Ultimately, however, the autonomy one can secure within a functionalist framework has principled limits: functional properties are defined in terms of causal roles that are occupied by *other* properties, so whatever they do, it's written into their metaphysical natures that something else is *really* doing it.[35] That something else is of course their basic physical realizers. This paper is part of an attempt to break free from functionalism. The traditional causal exclusion problem, in my view, arises not from causal closure, but from a conception of special science properties that comes with causal redundancy baked in. Geometric structure is higher-level and multiply realizable, but its realization does not consist in occupying a causal role. Ultimately, a conditioning role on the powers of basic physical properties stems either from the essential natures of those properties, or from the essential nature of spacetime, but the conditioning role itself, on some occasion, is *occupied by geometric structure* and not by its lower-level realizers on that occasion.

In closing, let me acknowledge a potential problem for the overall approach presented here.[36] One might object that geometric properties belong in higher-order explanatory contexts, in which one explains why a cause C explains a certain effect E, but not in first-order contexts in which one explains why E happens. Geometric properties might be needed if we want to explain why a system of multiple charges

[34] In Yates (2020) I tried to extend the simple account to neuroscience, through the idea that temporal patterns such as neural synchrony might be causally novel in the same way as the geometric structure of molecules. In a vast oversimplification, I likened the phase angle between oscillations in membrane potential to the spatial angles between atoms in a molecule. Still, I think there is something to the comparison: membrane potential is a vector quantity (the difference in potential between the inside and outside of the cell membrane) and the way in which distinct oscillating populations interact will be determined, inter alia, by the phase angle between their oscillations. This in turn will be at least partially explained by the way vectors compose.

[35] See Yates (2012) for more on the limits of functionalism. There I offer a grounding-theoretic account of the novelty of functional properties, according to which this novelty consists not in the powers they bestow, but in the distance, within a hierarchy of grounding relations, from which they bestow them. This account entails that all causal powers that special science properties bestow are ultimately bestowed by their basic physical realizers. For an attempt to circumvent these limitations of functionalism via what he calls *machresis*, see Gillett (2016).

[36] I thank Alastair Wilson for making me aware of the alternative that follows. See Hicks & Wilson (2021) for more on the distinction that follows between first- and higher-order explanations.

explains the acceleration of a charged particle, but they do not thereby belong in the first-order explanation of why that charged particle accelerated in the way that it did. What I herein regard as a single, unified multilevel explanation of the acceleration of the particle is really two distinct explanations, at different orders, with different explananda: a first-order explanation of the acceleration of the particle, and a second-order explanation of the first-order explanation.

It's tempting to note that the above objection is consistent with the claim that geometric properties have irreducible causal-explanatory roles, it's just that these roles now consist in explaining why certain first-order causal explanations hold. However, on reflection this doesn't seem to be much help when it comes to defending special science autonomy from the threat of causal exclusion. The causal exclusion problem is that there doesn't seem to be any causal work left for any properties other than those of basic physics, which threatens the *first*-order explanations that the special sciences provide of events such as molecular motions, metabolism, neural firings, and behaviours. Those events, exclusionists say, can in principle be explained in basic physical terms alone, so special sciences provide at most a more perspicuous way of explaining the same events. My strategy here has been to argue that causal explanations even in the simplest physics cases are irreducibly multilevel and that the complete cause of the relevant basic physical effects always involves interactions between properties at multiple levels. On the alternative just mooted, by contrast, the dynamics are wholly driven by basic physics, with only higher-order explanatory work left over for non-basic properties. This wouldn't matter so much were it not for the fact that the ambition of the special sciences doesn't seem to have anything much to do with explaining why first-order physical explainers explain. Rather, their central aim seems to be to provide distinctive *first*-order explanations of their target phenomena.

I don't know how to refute the claim that higher-level properties belong in higher-order explanations, but I can provide some additional motivation for my multilevel approach. If a physicist wants to *predict* the acceleration of a test charge placed in the field due to two or more charges, they first need to calculate the magnitude and direction of the resultant field at the point where the test charge is to be introduced. And to do so, I argued in Section 2, they must appeal to geometry. First-order predictions in such cases are simply not possible without appealing to higher-level geometric properties. But it would be very odd indeed to accept that higher-level properties are necessary for first-order predictions while at the same time denying that they feature in first-order *explanations*. Put differently, why is it necessary to appeal to geometric properties to predict the motion of a particle if such properties play no first-order causal role? If I am correct that geometric properties are among the manifestation conditions on basic physical powers, then it's easy to explain why they are necessary to predict how those powers manifest on some occasion. Such properties, on my view, are both predictively *and* explanatorily indispensable in first-order explanations of physical phenomena. Basic physical properties may bestow all the causal powers, but they don't occupy all the causal roles.[37]

References

Acuña, P. (2016). 'Minkowski Spacetime and Lorentz Invariance: The Cart and the Horse or Two Sides of a Single Coin?', *Studies in History and Philosophy of Modern Physics* 55, pp. 1–12.

Bartels, A. (1996). 'Modern Essentialism and the Problem of Individuation of Spacetime Points', Erkenntnis 45, pp. 25–43.

Brown, H., & Pooley, O. (2001). 'The Origin of the Spacetime Metric: Bell's "Lorentzian Pedagogy" and its Significance in General Relativity', in C. Callender & N. Huggett (eds.), *Physics Meets Philosophy at the Planck Scale: Contemporary Theories in Quantum Gravity*, Cambridge: Cambridge University Press (pp. 256–272).

Brown, H., & Pooley, O. (2006). 'Minkowski Space-time: A Glorious Non-entity', in D. Dieks (ed.), *The Ontology of Spacetime*, Amsterdam: Elsevier (pp. 67–89).

Brown, H. (2005). *Physical Relativity*. Oxford: Clarendon.

Crane, T., & Mellor, D. (1990). 'There is no Question of Physicalism', *Mind* 99, pp. 185–206.

Crook, S., & Gillett, C. (2001). 'Why Physics Alone Cannot Define the 'Physical', *Canadian Journal of Philosophy* 31, pp. 333–359.

Fodor, J. (1974). 'Special Sciences', *Synthese* 28, pp. 97–115.

Fodor, J. (1997). 'Special Sciences: Still Autonomous after all these Years', *Noûs* 31(S11), pp. 149–163.

Gillett, C. (2003). 'The Metaphysics of Realization, Multiple Realizability and the Special Sciences', *Journal of Philosophy* 100, pp. 591–603.

Gillett, C. (2016). *Reduction and Emergence in Science and Philosophy*. Cambridge University Press.

Heron, J., & Knox, E. (2019). 'On Constraints, Context and Spatiotemporal Explanation', *Philosophy and Phenomenological Research* 99, pp. 732–738.

Hicks, M., & Wilson, A. (2021). 'How Chance Explains', *Noûs* 57, pp. 290–315.

Jackson, F., & Pettit, P. (1990). 'Program Explanation: A General Perspective', *Analysis* 50, pp. 107–117.

Janssen, M. (2009). 'Drawing the Line between Kinematics and Dynamics in Special Relativity', *Studies in History and Philosophy of Modern Physics* 40, pp. 26–52.

Kim, J. (1992). 'Multiple Realization and the Metaphysics of Reduction', *Philosophy and Phenomenological Research* 52, pp. 1–26.

Kim, J. (1998). *Mind in a Physical World*. MIT Press.

Knox, E. (2019). 'Physical Relativity from a Functionalist Perspective', *Studies in History and Philosophy of Modern Physics* 67, pp. 118–124.

Lange, M. (2017). *Because Without Cause*. Oxford University Press.

LePore, E., & Loewer, B. (1987). 'Mind Matters', *Journal of Philosophy* 84, pp. 630–642.

List, C., & Menzies, P. (2010). 'The Causal Autonomy of the Special Sciences', in C. Macdonald & G. Macdonald (eds.), *Emergence in Mind*, Oxford University Press (pp. 108–128).

Maudlin, Tim (1989). 'The Essence of Spacetime', in A. Fine & J. Leplin (eds.), *Proceedings of the 1988 Biennial Meeting of the Philosophy of Science Association*, Vol. 2 (pp. 82–91).

[37] Versions of this paper were presented at a workshop in Birmingham on levels of explanation organized by Alastair Wilson's FramePhys project and at a joint LanCog-Complutense workshop in Madrid. My thanks to the audiences at both events for their comments and criticisms. I thank an anonymous referee for their comments. Special thanks to Amanda Bryant, Nicholas Jones, Katie Robertson, and Alastair Wilson for very helpful comments on earlier drafts. Based on research funded by the Fundação para a ciência e a tecnologia (awards: IF/01736/2014 and PTDC/FER-HFC/30665/2017).

Norton, J. (2019). 'The Hole Argument', in Edward N. Zalta (ed.), *The Stanford Encyclopaedia of Philosophy* (Summer 2019 Edition), https://plato.stanford.edu/archives/sum2019/entries/spacetime-holearg.

Papineau, D., & Spurrett, D. (1999). 'A Note on the Completeness of "Physics"', *Analysis* 59, pp. 29–32.

Pooley, O. (2006). 'Points, Particles and Structural Realism', in D. Rickles, S. French, & J. Saatsi (eds.), *The Structural Foundations of Quantum Gravit*y, Oxford University Press (pp. 83–120).

Shapiro, L. (2000). 'Multiple Realizations', *Journal of Philosophy* 97, pp. 635–654.

Shoemaker, S. (2001). 'Realization and Mental Causation', in C. Gillett & B. Loewer (eds.), *Proceedings of the Twentieth World Congress of Philosophy*, Cambridge University Press (pp. 23–33).

Wilson, J. (2006). 'On Characterising the Physical', *Philosophical Studies* 131, pp. 61–99.

Wilson, J. (2011). 'Non-Reductive Realization and the Powers-Based Subset Strategy', *The Monist* 94, pp. 121–154.

Wüthrich, C., Le Bihan, B., & Huggett, N. (eds.) (2021). *Philosophy Beyond Spacetime: Implications from Quantum Gravity*, Oxford University Press.

Yablo, S. (1992). 'Mental Causation', *Philosophical Review* 101, pp. 245–280.

Yates, D. (2009). 'Emergence, Downwards Causation, and the Completeness of Physics', *Philosophical Quarterly* 59, pp. 110–131.

Yates, D. (2012). 'Functionalism and the Metaphysics of Causal Exclusion', *Philosophers' Imprint* 12, pp. 1–25.

Yates, D. (2016). 'Demystifying Emergence', *Ergo* 3, pp. 809–841.

Yates, D. (2018). 'Inverse Functionalism and the Individuation of Powers', *Synthese* 195, pp. 4525–4550.

Yates, D. (2020). 'Neural Synchrony and the Causal Efficacy of Consciousness', *Topoi* 39, pp. 1057–1072.

Yates, D. (forthcoming). 'Hylomorphism, or Something Near Enough', in D. Yates & A. Bryant (eds.) *Rethinking Emergence*, Oxford University Press.

PART III

LEVELS OF EXPLANATION IN HIGHER-LEVEL SCIENCES

7
Explanatory Levels in Living Organisms

William Bechtel

1 Introduction

The notion of levels is used polysemously in science, especially in the life sciences (see Craver's field guide to levels in his 2007, Ch. 5). On some conceptions of levels (e.g., a conception on which levels are differentiated by the sizes of the entities), levels span the universe, supporting a stratified representation. This is not true, however, of two conceptions of levels that figure centrally in biology: mechanistic levels and levels of control. Each identifies levels in local contexts, but fails to generate stratified levels that extend across biological phenomena. This, however, does not impair the usefulness of thinking of mechanisms or control in biology in terms of levels when the resulting conception of level is appropriately restricted.

Mechanistic levels have received considerable philosophical attention with the emergence of new mechanistic accounts of explanation (Machamer, Darden, & Craver, 2000; Bechtel & Abrahamsen, 2005; Glennan, 2017). The new mechanists characterize biologists as appealing to mechanisms to explain phenomena—the explanation identifies a mechanism that is taken to be responsible for a given phenomenon, decomposes it into parts or entities, each performing operations or activities, and then shows how, when appropriately organized, these components produce the phenomenon. The components of mechanisms are then represented as at a lower level than the mechanism itself. Insofar as the components constituting the mechanism are often themselves mechanisms consisting of their own parts carrying out activities, biologists often iterate this process until they reach what Machamer et al. refer to as a bottom-out level of entities and activities, for which they do not seek an explanation. Each level of decomposition of a mechanism into its parts results in a lower level of mechanisms. A given inquiry can generate a hierarchy of levels. As I explain below, these hierarchies are only defined locally—they don't define a general stratification of entities.

As biologists investigate mechanisms, they often discover that the mechanisms they were investigating only operate under some circumstances. Recognizing this, some biologists turn their attention to processes outside mechanisms that control their operation. For the most part, new mechanists have not focused on the control of mechanisms. Accounts of control have, however, become increasingly prominent in biology. When biologists identify the processes that control mechanisms,

William Bechtel, *Explanatory Levels in Living Organisms* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © William Bechtel 2024. DOI: 10.1093/oso/9780192862945.003.0008

they often characterize the processes doing the controlling as themselves mechanisms. To provide clarity, I will refer to such mechanisms as *control mechanisms*. Like other mechanisms, biologists decompose control mechanisms in order to explain their operation. But they also seek to understand how they operate to control other mechanisms in response to conditions arising within the organism or its environment. In doing so, they often represent control mechanisms as at a higher level than the mechanisms they control.

To understand how control mechanisms regulate other mechanisms, it will help to supplement the framework of the new mechanists with an ingredient that is required for any mechanism to operate—free energy (the energy available to do work). The entities constituting mechanisms constrain the flow of free energy so that particular work is performed (Kauffman, 2000). The relevant notion of constraint is drawn from classical mechanics (Sklar, 2013), where it is characterized in terms of limitations imposed on the degrees of freedom of movement of each particle when two or more are bound together. As developed by Hooker (2013), constraints are both limiting and enabling. By limiting the directions in which a liquid can flow, a pipe directs it to reach a location it otherwise wouldn't. On this view, the activities of mechanisms just are the work performed as free energy is constrained (Winning & Bechtel, 2018). To understand how a mechanism can then be controlled, one must make a distinction between constraints. Many constraints in a mechanism are fixed on the timescale at which the operation of the mechanism is being characterized. But others are flexible. Changes in them alter what the mechanism does. Control mechanisms alter the behavior of other mechanisms by operating on their flexible constraints, thereby changing their behavior (Winning & Bechtel, 2018).

Control mechanisms, like all mechanisms, perform their work as a result of how their parts constrain the flow of free energy through them. For them to provide control that responds to conditions within the organism or its environment, the flexible constraints within the control mechanism must be responsive to those conditions. As a result of its constraints being configured to respond to these conditions, control mechanisms can be characterized as measuring the value of variables representing conditions within the organism or its environment and altering the activities of other mechanisms based on those measurements.

Insofar as control mechanisms are often operated on by other control mechanisms, biologists often refer to levels of control. These levels are, though, importantly different from the compositional levels frequently invoked by the new mechanists. Accordingly, in Sections 2 and 3 I will further explicate each and illustrate how each is invoked in understanding a specific biological phenomenon—movement generated by skeletal muscles in animals. In each case I will argue that the locally characterized notion of levels is adequate for the role for which it was developed without leading to a general stratification of levels. Philosophers can analyze biologists' practices of advancing mechanistic explanation and of accounting for control in organisms in terms of locally characterized notions of level. I conclude in Section 4 that to understand mechanisms and how they are controlled, neither biologists nor philosophers of science analyzing biology require a general stratification of biological entities into levels.

2 Locally Defined Levels of Mechanisms

Mechanistic levels are defined with respect to the mechanisms biologists identify as the loci of biological phenomena. What is characteristic of mechanistic explanation is the appeal to the components of the mechanism to explain a phenomenon. Components may be identified either structurally in terms of component entities or functionally in terms of the component activities that together produce the phenomenon. The tools for decomposing a mechanism into entities and into activities are different, and at different stages of inquiry researchers may only have one available. A general aspiration is to be able to map activities onto entities (Bechtel & Richardson, 1993/2010, refer to this as localization). These component entities and activities are treated as the denizens of the level below the mechanism. These components can in turn be decomposed into their entities and activities, with these subcomponents occupying yet a lower level (Figure 7.1). Starting from a given mechanism, one can also identify a mechanism of which it is a component. It and the other components with which it interacts in producing the phenomenon associated with that mechanism are the denizens of this level. The componential relationship between components and mechanisms is what gives rise to a hierarchy of levels. Before discussing this further, I provide an example of an explanation that spans multiple mechanistic levels.

A notable feature of animals, celebrated by Aristotle, is their ability to move their limbs or segments of their body in a manner that enables them to propel themselves through space.[1] Locomotion is a phenomenon to be explained. The first level of explanation appeals to muscles—tissues consisting of contractile components known as fascicles, surrounded by a perimysium that groups fascicles into bundles, and tendons that attach these bundles to the skeleton. These need to be organized in such a manner that the contraction of the fascicles results in the movement of the parts of the animal's body. To explain the contraction of fascicles researchers decompose them into muscle fibers. Muscle fibers are unusual cells that contain multiple nuclei; they form during development from the merger of more traditional cells with single nuclei. At this level, microscopic visualization of muscles reveals that they are longitudinally divided by Z lines into units known as

[1] A further form of movement in animals is the movement of internal organs—notably, the heart, lungs, and digestive tract. A similar explanation can be provided for this movement as for locomotion, but I will focus on locomotion and the skeletal muscles responsible for it.

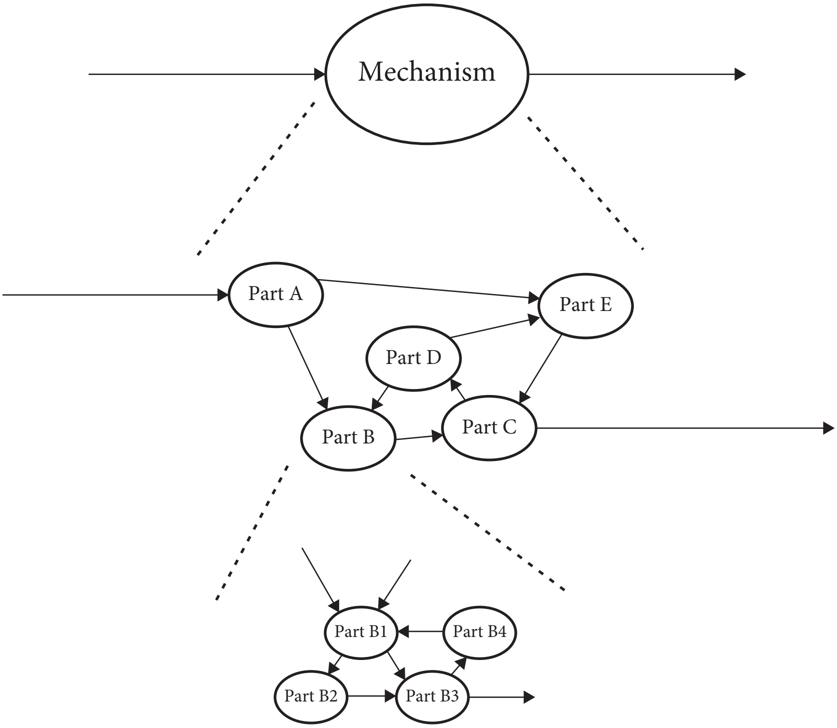

Figure 7.1 The mechanism represented by the top oval is decomposed into component parts A, B, C, D, shown in the middle that together define a lower level. At the bottom Part B is shown as further decomposed into B1, B2, B3, B4, which constitute a yet lower level.

sarcomeres. Within each sarcomere is one dark or anisotropic (A) band and parts of two light or isotropic (I) bands, each of which extends across the Z line into the next sarcomere (Figure 7.2). During contraction, the I bands shorten and the distance between Z lines is reduced. The shortening of the I bands offers an explanation of the contraction of muscle fiber.

The shortening of the I bands itself calls out for explanation. Developing such an explanation required researchers to decompose the A and I bands into molecules and to determine what those molecules do. In the late 1940s Szent-Györgyi differentiated two fibril proteins that interact in muscle contraction: actin and myosin. I bands were determined to consist of actin and A bands of overlapping actin and myosin fibrils. With the development of new techniques of microscopy, A. F. Huxley and R. Niedergerke (1954) and H. E. Huxley and J. Hanson (1954) demonstrated that when muscles contract, actin fibrils are drawn along myosin fibrils. In subsequent research, H. E. Huxley (1969) advanced the now generally accepted swinging crossbridge explanation in which a part of the myosin molecule

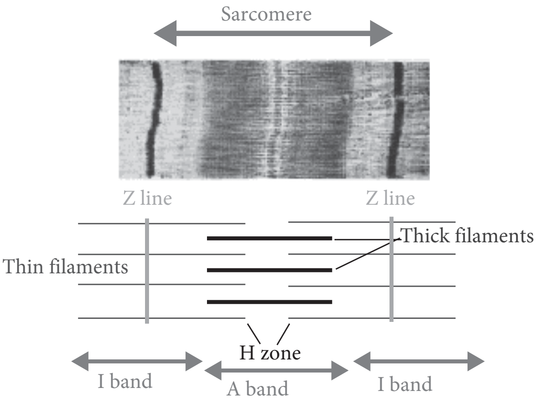

Figure 7.2 In muscle contraction, actin and myosin filaments slide along each other, resulting in reducing the I-Band and pulling the Z-discs closer to each other. Figure by Sameerb, released for any purpose on Wikicommons.

referred to as the head successively binds to and releases from actin. When bound it executes a powerstroke that pulls the actin alongside itself; when unbound it straightens out before binding again at a location further along the actin filament. The powerstroke requires a source of free energy, which is provided by hydrolyzing adenosine triphosphate (ATP) (removing the gamma phosphate group from ATP, yielding adenosine diphosphate (ADP)). Lymn and Taylor (1971) described a cycle in which myosin binds ATP (Figure 7.3, panel A), hydrolyzes it (panel B), and then again attaches itself to actin (panel C). Finally, myosin carries out the powerstroke (panel D).

In performing these activities, myosin undergoes a series of conformation changes—changes in the spatial arrangement of its constituent atoms. The overall movement of the head can be observed in electron micrographs, but the details of how atoms are moved within myosin required employing protein crystallography. With it, Rayment et al. (1993) revealed a binding site for ATP at the opposite end of a β-sheet from an actin-binding region, leading to the hypothesis that the altered arrangement generated by the removal of the gamma phosphate from ATP altered the β-sheet, thereby altering the actin binding site in a manner that enabled binding to actin. The images also revealed a long tail, consisting of an α helix, which has the appearance of a lever arm. Comparison of images corresponding to myosin binding ATP and ADP revealed a change in the position of the lever arm, supporting the hypothesis that the energy liberated in hydrolysis is stored in the conformation change of the lever arm until it is released in the powerstroke (Rayment, Smith, & Yount, 1996).

Each level in this mechanistic scenario characterizes components of the entities at the higher level. Fascicles and tendons are components of muscles. Muscle fibers

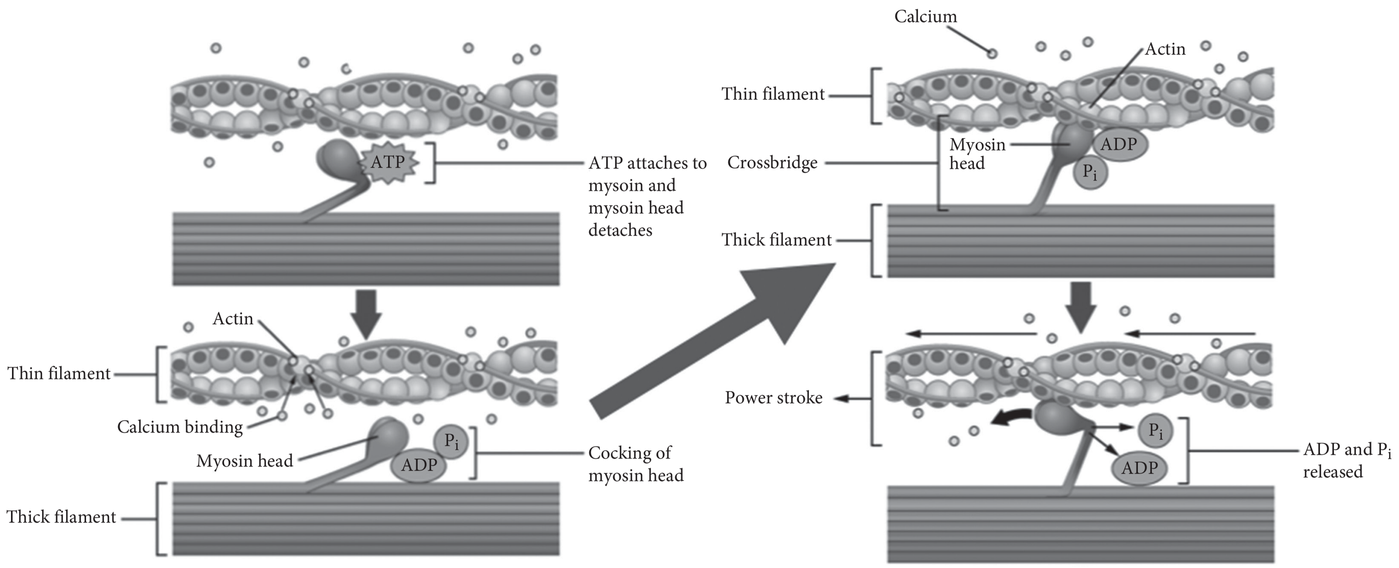

Figure 7.3 Stages in the Lymn-Taylor cycle through coupled to the states of the myosin head. Reprinted from J. G. Betts, et al. (2013), figure 10.11, released under CC BY 4.0 license by OpenStax, https://cnx.org/contents/FPtK1zmh@8.25:fEI3C8Ot@10/Preface. For enquiries concerning use outside the scope of the licence terms, please contact the rights holder.

(cells) are components of fascicles. Actin and myosin molecules are, in turn, components of these muscle fibers, and the ATP- and actin-binding regions and the lever arm are components of myosin. The entities at each level perform different activities. Muscles, as wholes, enable movement of limbs through the exertion of force by fascicles on tendons—fascicles contract, myosin filaments pull against actin filaments, and, within myosin, hydrolysis at the ATP-binding site generates torque on the lever arm. At each level the components need to be properly organized; fascicles need to be attached properly to tendons, muscle fibers in fascicles need to be appropriately oriented vis à vis each other, heads of myosin need to be able to form crossbridges with actin and exert force on it, and within myosin the binding pockets for ATP and myosin as well as the lever arm need to be appropriately situated vis à vis each other so that force exerted within one alters the conformation of the others.

It is the decomposition of mechanisms into component mechanisms that gives rise to a hierarchy of levels. Researchers decompose the mechanism taken to be responsible for the phenomenon into components that act together to produce the phenomenon. These components are at a lower level than the mechanism as a whole. Researchers can then treat each component as a mechanism and decompose it. The entities identified in that decomposition constitute the next-lower level.

The mechanistic conception of levels differs from conceptions of levels such as that proposed by Churchland and Sejnowski (1988), on which levels are defined in terms of the size of their constituents. Other than requiring that components be smaller than the whole mechanism, the mechanistic account is agnostic about size.[2] What renders entities at the same level is that they are taken to interact in a mechanism to produce a phenomenon. If molecules are taken to interact with membranes or even with cells in generating a phenomenon, they together constitute a mechanistic level in that decomposition despite the differences in their sizes.

Critically, mechanistic levels are decomposition relative; if researchers employ a different decomposition, they may place different entities at the same level. For example, if one researcher adopts the decomposition shown in Figure 7.4 (left panel), she will place C, D, and E the same level as A and B. If another researcher adopts the decomposition shown in Figure 7.4 (right panel), she will place F at the same level as A and B and C, D, and E at a lower level than F. A more general consequence of this is that the mechanistic account of levels does not specify whether subcomponents of two different components of a mechanism are at the same level. If Figure 7.1 had shown the decomposition of another component shown in the middle layer, the account would not determine whether the subcomponents in the two decompositions are at the same level since they are not part of the same

[2] The mechanistic account is not unique in this respect. See, for example, List's discussion of supervening levels in Chapter 1.

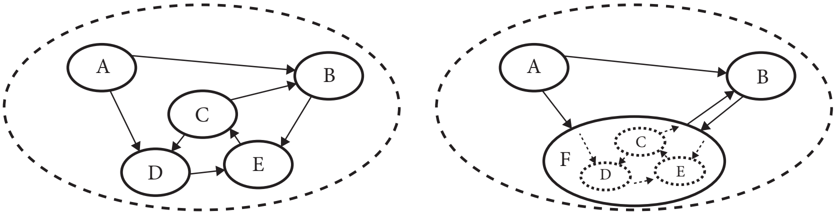

Figure 7.4 Whether C, D, and E belong to the same level as A and B depends on whether they are viewed as components directly interacting with A and B or as themselves components of an entity that interacts with A and B.

mechanism. All one could say is that each is at a lower level than the component of which it is a subcomponents. More generally, assignment of entities to levels depends upon how researchers have decomposed mechanisms. Depending on which decomposition is employed, two entities might or might not be situated at the same mechanistic level.

Despite identifying levels only locally, the mechanistic conception of levels suffices for scientists in discussing the explanations they propose. It captures what they intend when they characterize mechanistic explanations as reductionistic—they appeal to entities or activities at lower levels to explain a given phenomenon. This conception of reduction differs from philosophical accounts in which there is a lowest level from which accounts of higher-level phenomena can be derived. The mechanistic conception of levels does not offer a means of characterizing a lowest level from which different biological phenomena are constructed, but only levels defined with respect to how a given mechanism is decomposed. It is also worth noting that in addition to appealing to components, mechanistic explanations appeal to how the components are organized—they do not propose to account for phenomena solely in terms of lower-level components. Accordingly, in developing reductionistic mechanistic explanations there is little motivation for identifying a comprehensive lowest level.

Although restricted, the mechanistic account of levels suffices to understand how scientists discuss both bottom-up and top-down causation without impugning to them a commitment to over-determination of the sort criticized by Kim (1998). What is required is to limit causal characterizations to interactions between entities at the same mechanistic level. Causal interactions are what enable entities to work together to produce a phenomenon; accordingly, the arrows in Figures 7.1 and 7.4 are all between entities at the same level. The relation between entities or activities at different levels is not causal but componential: the mechanism consists of its component mechanisms. These do not cause its behavior, nor does the whole mechanism cause the activities of its components. When a cause acts on a mechanism, the effects are manifest at both lower and higher levels—in the components of the mechanism and in any larger mechanism of which the mechanism is a component. A cause cannot produce a change in a mechanism without changing at least one of is components, and one cannot change a component without changing in some manner the mechanism of which it is a component. Craver and Bechtel (2007) thus proposed that we understand scientists' appeals to both bottom-up and top-down causation in terms of a constitutive relation between levels and causal interaction within levels. On this rendering, talk of top-down causation is no more mysterious than talk of bottom-up causation. It simply recognizes that when some causal process affects a mechanism, its effects will show up in the activities of one or more of its components. One doesn't need, in addition, to characterize the whole as acting on its components causally.

While the conception of mechanistic levels is clearly limited, it suffices for understanding mechanistic explanations in biology and the inquiries that contribute to their development. A mechanistic explanation shows how, given the coordinated activity of components, a particular phenomenon is produced. It provides a conceptual framework to integrate into one account activities occurring within a mechanism and how the mechanism interacts with entities at its same level. Key to the account of levels is the manner in which researchers decompose a mechanism into components and recompose the mechanism from its constituents. The resulting mechanistic explanation is inherently a multilevel account—it appeals to both components that are denizens of a lower level and the manner in which they are organized into the mechanism which is observed interacting with other mechanisms at the higher level.

The account I have advanced so far adheres to the formulations of the new mechanists that view mechanisms as decomposable into entities and activities. In Section 1 I introduced an alternative account in which mechanisms constrain the flow of free energy to perform work. Before leaving the mechanistic conception of levels, I will briefly consider how the mechanistic account of levels appears under this alternative conception of mechanism. The difference between the accounts can be seen by considering the notion of activity. Machamer et al. (2000) argue for a dualism of entities and activities—activities are not explicable in terms of static entities and, as the term suggests, constitute active doings in the world. On their account, any explanation of an activity must itself appeal to other activities. Machamer et al. chose the term *activity* to emphasize that the components of mechanisms are active—they do things—but they offer no explanation of what makes activities active. The revisionist account in terms of free energy and constraints is also dualistic, but the notion of free energy is grounded in the fundamental understanding of how the universe works provided by thermodynamics. Although according to the first law total energy in a system is conserved, according to the second law the energy available for work, free energy, is continually lost as heat. Work can be performed even as the system progresses toward equilibrium by constraining some of the free energy as it dissipates. Work is then manifest in various activities. In particular, the activities of organisms result from how they constrain free energy procured from their environment. Organisms capture free energy in molecules in individual cells. Sugars and fats, which animals procure in their diets, provide a source of free energy, which they convert into such currency as ATP, which is used in carrying out individual activities such as muscle contraction.

On the revisionist account, any activity performed by any mechanism results from free energy and constraints. Explanation involves showing how the free energy available to the mechanism is constrained. As researchers decompose the mechanism to explain how it works, they decompose the system of constraints into entities taken to be at a lower level. They do not, however, decompose the free energy—rather, they identify where it figures in the analysis of the mechanism at that level. After the discovery of ATP in 1930 and the recognition that it, not heat, provided the free energy for biology reactions through the hydrolytic breaking of the bond to the γ-phosphate, muscle researchers viewed it as providing the free energy for the observed shortening of the I-bands in muscle contraction. After H. E. Huxley decomposed that activity into the activities involving the myosin head, Lymn and Taylor linked these activities to the steps in ATP hydrolysis. Once it was possible to decompose myosin into component parts of the molecule, researchers could identify the site of ATP binding and propose how the conformation change resulting from hydrolysis exerted force to move the lever arm. At each level that mechanism is decomposed, but the same free energy is identified at each level.

With the determination of the molecular structure of myosin and how it constrains free energy released in ATP hydrolysis to move a lever arm, the mechanistic explanation of muscle contraction has likely bottomed out (Bechtel & Bollhagen, 2021). This account specifies both the form free energy takes and how it is transferred to produce the phenomenon. But there are two things to note. First, researchers could choose to terminate their explanations at higher levels. Since they could not analyze the molecular structure of myosin, out of necessity H. E. Huxley and Lymn and Taylor offered explanations in terms of the swinging crossbridge and ATP hydrolysis without going to a lower level. Moreover, it appears their explanation correctly described the operation of the mechanism and accounted for the phenomenon of muscle contraction. The further decomposition was not required to produce the critical evidence that the explanation was correct—that resulted from the single-molecule assays created by Finer, Simmons, and Spudich (1994) that enabled the measurement of the force generated and movement produced by a single myosin molecule hydrolyzing ATP. In general, researchers can bottom out their inquiry when they have identified constraints on free energy that account for the phenomenon in which they are interested. Second, there is not a single mechanistic level at which the full explanation of muscle contraction is provided. The account of force generation within myosin can be carried out at the level of molecular structure. But that falsely suggests that there is a univocal level at which the whole explanation of muscle contraction is presented. That isn't the case—as researchers decomposed the original mechanism, they identified the pertinent constraints at each iteration of decomposing the mechanism.

In this section, I have shown how decomposing mechanisms into components provides biologists a strategy for explaining biological phenomena. Decomposition identifies entities at a lower level than the mechanism itself. But such levels are only locally defined. Biologists start from phenomena and work down. The entities that interact in the mechanism to produce the phenomenon are viewed as situated at a common lower level. But this does not specify how these entities are related to those identified in explaining other phenomena. Moreover, different scientists can decompose the same mechanism in different ways (e.g., one treating as a component what another treats as a cluster of components). Even though individual explanations bottom out, the practice does not reveal a bottom level that spans biology. Levels defined relative to a particular decomposition nonetheless suffice for the purposes of developing explanations in biology. They are all philosophers of science need to characterize explanations in biology.

3 Locally Defined Levels of Control

I turn now to control processes and the concept of levels generated by control mechanisms operating on other mechanisms. The need for control mechanisms in biology is evident in the example of muscle contraction. Without control, muscles would contract any time free energy is available and would continue to do so until free energy is exhausted. This would result in rigor mortis. To avoid rigor mortis, muscles must be stopped from contracting even when free energy is still available. More generally, for an organism to maintain itself, its various muscles must contract under appropriate conditions, and not otherwise.

The key to control is that some of the constraints within a mechanism are flexible. As a result, other mechanisms can act on these constraints and thereby change (stop, start, redirect, or modulate) how the controlled mechanism works. Switches on human-made machines are exemplars of flexible constraints—a user can turn on or turn off the machine by flipping the switch. The human user turns on a machine when he or she detects conditions in which the machine's activity is desired. The same is true of a control mechanism—it acts on and changes a flexible constraint when it detects conditions to which it is equipped to respond. The essential feature of a control mechanism is that when it detects conditions to which it is designed (by humans or evolution) to respond, it initiates action on a flexible constraint of the mechanism it controls.

Control processes operate on the muscle mechanism discussed above. For myosin to exert force on actin, it needs to bind to it. By default, however, another molecule, the troponin complex, binds to the site, blocking myosin's access to it.

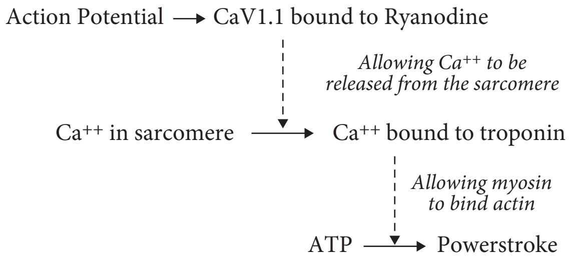
Action Potential
CaV1.1 bound to Ryanodine
Allowing Ca++ to be released from the sarcomere
Ca++ in sarcomere
Ca++ bound to troponin
Allowing myosin to bind actin
ATP
Powerstroke

Figure 7.5 Two levels of control mechanisms acting on the mechanism generating the power-stroke of myosin on actin.

To enable contractions, troponin must be removed from the binding site. The conformation of troponin is flexible, and only in some conformations can it bind actin. Binding calcium ions (Ca^{++}) at another site on troponin forces troponin into a conformation in which it cannot bind actin. This makes it possible for myosin to bind to actin, allowing for the activities discussed in the previous section to proceed. As a result, the availability of Ca^{++} controls whether myosin binds to actin and the muscle contracts. Troponin detects the presence of Ca^{++} and allows myosin to act on actin only when it is available.

Since a control mechanism operates on constraints in the controlled mechanism and thereby changes what it does, control mechanisms are commonly characterized as at a higher level than those they control. Thus, in Figure 7.5, the control activity of Ca^{++} ions is shown as at a higher level than the action of ATP in generating a powerstroke. To indicate that the action of Ca^{++} bound to troponin causes the change in constraints that allows ATP to bring about the powerstroke, the relation is shown with an arrow. The fact that the arrow is dashed simply reflects that this causal process is a control process operating on a constraint in the controlled mechanism.

If a control mechanism contains a flexible constraint, it can be acted on by another control mechanism, adding another level of control mechanism. In the case of muscle, a further mechanism controls the availability of Ca^{++}. Generally, Ca^{++} is kept sequestered in the sarcoplasmic reticulum—a membrane enclosed structure in the cell—and is only released when channels that are part of ryanodine receptors are open. Normally they are closed. They are only opened when another molecule, CaV1.1, binds to ryanodine (RyR1 in Figure 7.6). In its default conformation, CaV1.1 cannot bind ryanodine. A further condition is required to alter its conformation—an action potential in a motor neuron enervating the transverse tubule. CaV1.1 is a voltage-dependent Ca^{++} channel situated in the membrane of transverse tubules. The processes that make it voltage dependent enable CaV1.1 to detect action potentials. Accordingly, as shown in Figure 7.6, when an action potential causes CaV1.1 to bind ryanodine, that opens a channel that allows Ca^{++} to

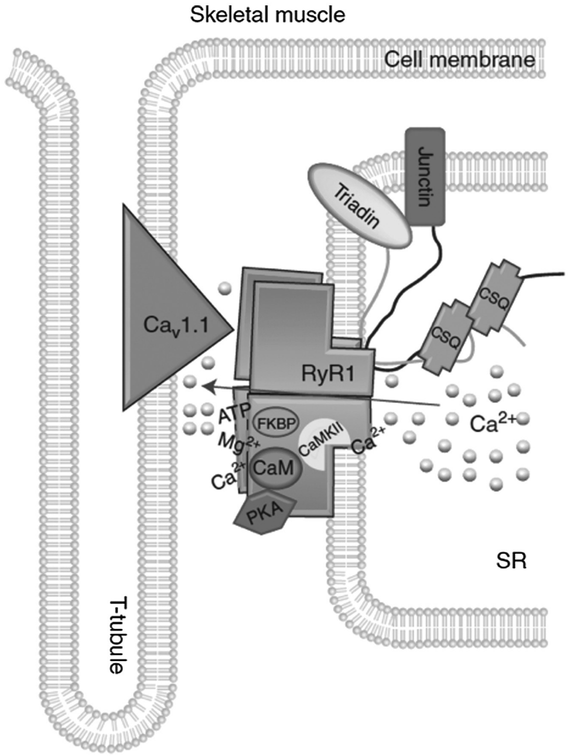

Figure 7.6 Release of Ca^{++} from the sarcoplasmic reticulum (SR) by the ryanodine (RyR1) receptor is regulated not only by CaV1.1 but by many additional molecules. CaV1.1 detects action potentials from neurons enervating the transverse tubule. From Lanner, Georgiou, Joshi, and Hamilton (2010). This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

be released from the sarcomere and bind with troponin. CaV1.1's response to an action potential constitutes a second level of control.

The account of muscle control I have presented so far vastly oversimplifies the control process. Figure 7.6, for example, shows numerous other molecules associated with the ryanodine receptor: protein kinase A (PKA), Ca^{++}-calmodulin-dependent protein kinase II (CaMKII), FK506-binding proteins, calmodulin (CaM), calsequestrin (CSQ), triadin, junctin, as well as various small molecules such as ATP, Mg^{++}, and Ca^{++} itself. Several of these are assumed to be involved in detecting or measuring other conditions and modulating the behavior of the ryanodine receptor. In many cases researchers have not yet determined what role these molecules play. One case that is understood involves PKA. A very common way of altering the behavior of a protein is to cause a phosphate ion (PO_4^{3-}) to bind to one of its amino acids, a process known as phosphorylation. This reaction is typically catalyzed by a molecule known as a kinase. PKA is a kinase that when activated phosphorylates the ryanodine receptor, making it more active. PKA is thus a second flexible constraint that allows for higher-level control. PKA is activated by a sequence of chemical reactions initiated by the binding of the neurotransmitter noradrenaline to β-adrenergic receptors on the cell membrane. Noradrenaline is released by neurons in the sympathetic nervous system in conditions that initiate a fight-or-flight response. The phosphorylation of ryanodine receptors by PKA results in faster muscle contraction, enabling the needed response. The action of noradrenaline on PKA and hence on ryanodine receptors is independent of that of action potentials in motor neurons.

The fact that two, and likely more, mechanisms are involved in setting constraints in the ryanodine receptor highlights a common feature of control in biology—individual mechanisms are often controlled by multiple control mechanisms. This differs from how control is realized in human institutions. Typically, a worker only reports to one supervisor. In a case where a worker answers to two or more, she may receive conflicting commands. If no external source is available to adjudicate the conflict, she has to decide which supervisor to obey. The same is true in biology—the controlled mechanism determines the response when multiple higher-level control mechanisms act on constraints and direct the operation of the controlled mechanism in different ways.

Allowing only one controller to operate on a controlled mechanism is one feature of hierarchical organization. When only one control mechanism acts on a controlled mechanism but a control mechanism is able to act on multiple controlled mechanisms, there are fewer controllers at higher levels in the hierarchy. Thus, the organization of companies is often represented as a pyramid with fewer controllers at each level until at the highest level there is one CEO. This structural organization is accompanied by functional relations in which information flows upward to the CEO and directives from the CEO are passed down to the workers who execute them. The organization of biological control mechanisms deviates in many ways from that of a hierarchical pyramid. As a result of multiple control mechanisms acting on a single controlled mechanism, there can be, and generally are, more controllers at higher levels. The pyramid is often inverted: rather than a single control mechanism serving as the central executive, one often finds multiple control mechanisms acting on individual mechanisms. McCulloch (1945) introduced the term *heterarchy* to describe human preference rankings that violate hierarchical ordering, and Pattee (1991) extended the term to control in biology.

The assumption that control mechanisms are organized hierarchically is deeply entrenched. The brain is thought to be the controller of the body, and the brain itself is often presented as hierarchically organized. The cerebral context is taken to be the highest-level control system, with regions of the frontal cortex constituting the central executive. Embracing this assumption, brain research often starts by investigating the neocortex, examining how it receives information from what are regarded as lower levels within the brain and sending motor commands to those regions. There is no question that frontal and prefrontal areas play important roles in regulating motor behavior. Experimental manipulations of specific neurons in the premotor cortex, for example, can elicit or inhibit activity of corresponding muscles. But this does not show that all other control processes operative in an organism are subservient to those in frontal and prefrontal regions of the neocortex and that the overall organization is hierarchical.

One can appreciate how heterarchical control of muscle is if one starts from the muscles themselves, as I did above, and continues to move up levels of control. The action potentials which the CaV1.1 receptors detect are produced in neurons that originate in the spinal cord, where they either belong to or are activated by what are known as *central pattern generators* (CPGs). These small networks of neurons generate rhythmic patterns that drive activity in the muscles to which they are connected. CPGs not only generate rhythms but integrate inputs from multiple sources. One source is somatosensory feedback from the muscles themselves. By taking this into account, CPGs are able to adjust the signals they send to muscles in response to how the muscle responds to efferent signals. Another source is other CGPs—receiving inputs from other CPGs enervating nearby muscles enables them to generate coordinated contractions. Yet another source of inputs to CPGs is the midbrain mesencephalic locomotor region (MLR). Stimulation of neurons in the MLR can elicit coordinated sequences of muscle activity such as is required for walking and running.

Researchers committed to control being hierarchical present these spinal cord and midbrain areas as just implementing commands issued from the neocortex. That this is not the case was shown by studies in which the cerebral cortex was removed in infant cats. Bjursten, Norrsell, and Norrsell (1976) found that these cats were able to perform their normal daily activities—eating, moving about, cleaning themselves—and lived autonomously for years in the protected environment of a laboratory. If subcortical regions such as the basal ganglia, hypothalamus, and thalamus are destroyed as well, but the MLR is preserved, cats can continue to produce motor responses when the MLR is stimulated but do not initiate motor activity (Shik & Orlovsky, 1976). This suggests that some of these subcortical areas are part of the mechanism that exercises control to initiate motor activities. Studies of these areas have revealed some of the conditions to which neurons in them are responsive, but do not support the idea that any one of them acts as the top-level executive. The hypothalamus contains multiple nuclei (interconnected neurons) that are responsive to conditions within the organism, such as nutritional status or sleep deficits, and initiate and terminate actions such as feeding and sleeping when appropriate conditions are registered (Leng, 2018).[3] One of the reasons some

[3] Areas in the brainstem such as the nucleus of the solitary tract also register these conditions and in some cases initiate activity even without the hypothalamus.

think a central executive is required is that otherwise distributed controllers such as those in the hypothalamus would initiate incompatible actions—eating while sleeping. Selecting among alternative actions is an important control activity, but it is performed by the nuclei constituting the basal ganglia. By default, the output nuclei inhibit other brain areas; other nuclei of the basal ganglia enable a competition between inputs from other brain areas, acting to remove the inhibition from the area sending the stronger input (Redgrave, Vautrelle, & Reynolds, 2011). Subcortical neural processes exhibit a similar heterarchical organization we observed within muscle cells—multiple control processes interacting among themselves and each acting upon the CPGs that send signals to muscles.

A final illustration of the heterarchical organization of control structures in animals is provided by neurotransmitters that serve as neuromodulators (Katz, 1999). These operate in a very different manner than the more familiar process in which neurotransmitters that are released at the synapses act upon ionotropic receptors in postsynaptic neurons, initiating the flow of current along the membrane of the postsynaptic cell. Neuromodulators can be released at locations other than the synapse, including the cell body and along both dendrites and axons. From these locations they diffuse through the extracellular matrix until they reach a cell which has a metabotropic receptor to which they can bind. When the transmitter binds a metabotropic receptor, it initiates a cascade of reactions, modifying the metabolism of the recipient cell. This can result in new gene expression that, among other things, alters the responses of receptors at synapses for the more familiar neurotransmitters that act on ionotropic receptors. Both the diffusion of neuromodulators and the action on metabotropic receptors occur over a much longer time span than the operation of a transmitter acting on an ionotropic receptor, and their effects can endure over extended periods. Hence, they are characterized as modulating the behavior of other neural circuits, but they might better be conceptualized as setting the agendas for synaptic processing by neurons in regions to which they project (Bechtel, 2022).

The action of noradrenaline, discussed above as acting on ryanodine receptors in flight-or-flight situations, resulting in faster muscle contraction, is an example of the control activity of neuromodulators. The process leading to the release of noradrenaline is initiated by activity in the paraventricular nucleus of the hypothalamus, which stimulates the pituitary gland to release corticotropin into the bloodstream. When corticotropin is detected at the adrenal gland, it releases noradrenaline. This complex sequence of events is similar to those involving other neuromodulators such as dopamine, serotonin, as well as dozens of peptide neurotransmitters. All of them initiate widespread slow responses that are long lasting. (See Hills et al., 2015 for an illuminating discussion of the role of dopamine in regulating search processes, whether it be physical search of the enviornment or mental search.)

This brief examination of control of skeletal muscles reveals important aspects of how the notion of levels figures in discussion of control in biology and how that notion of levels differs from mechanistic levels. The motivation for treating control mechanisms at a higher level is illustrated in Figure 7.5: control mechanisms act on and modify the causal behavior of the mechanisms they control. The relation between levels is causal, not compositional. The neat ordering of levels in Figure 7.5, however, does not reflect the complexity of control illustrated in the case of muscles. Rather than just one control mechanism operating on a controlled mechanism, there are often many. Figure 7.7 shows three control mechanisms, a, b, and c, operating on flexible constraints in myosin, thereby controlling muscle contraction. One consequence noted above of having multiple control mechanisms operating on the same controlled mechanism is that the various controllers can conflict, one changing constraints in the controlled mechanism so that it operates in one way and another changing constraints so that it operates differently. Unless there is another mechanism adjudicating the conflict, the controlled mechanism will produce whatever behavior it does with both modifications to its constraints.

Figure 7.7 portrays control mechanisms at three different levels. But it also illustrates ways in which the representation of distinct levels of control can be compromised. For example, control mechanisms represented at different levels (f and h) can both act on the same controlled mechanism. In the control of muscles, the MLR receives control inputs from both the midbrain and from the neocortex. There are good reasons to view the neocortex as at a higher control level than the various midbrain regions, but it also sends outputs directly to mechanisms more than one level lower. In addition, in Figure 7.7 there are two upward arrows terminating on the detectors of control mechanisms (d and h) at higher levels. This reflects the fact that the conditions a control mechanism detects may include states of the mechanism that it is controlling. I noted above that CPGs typically receive feedback from the muscles they are regulating. Human designers often implement such feedback in machines—the governor Watt designed for the

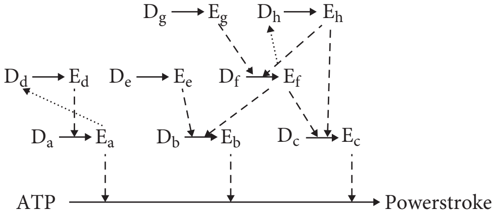

Figure 7.7 Multiple control mechanisms, schematized by a solid arrow from the detector (D) to the effector (E) and a dashed arrow to the mechanism being controlled. There are upward two dotted arrows indicating detectors that are responding to conditions in mechanisms at lower levels.

steam engine detects when the speed of the engine is too fast or too slow and acts on a valve (a flexible constraint) to modify the flow of steam. Where there is feedback, causal processes operate both upward and downward.

As with mechanistic levels, differentiating levels of control works well locally—for any controlled mechanism, one can conceptualize the control mechanism as at a higher level. But the process of evolution is not constrained to respect hierarchical organization. Control mechanisms can be added in organisms opportunistically as long as the consequences are not fatal. Likewise, new outputs from one control mechanism to other mechanisms are easy to add. As a result, when biologists try to represent the multitude of control processes operative in living organisms, they often abandon layered diagrams such as Figure 7.7 and develop network diagrams in which each entity figuring in the control process is represented as a node and its various effects on other nodes is represented as an edge (Bich & Bechtel, 2022, this practice in representing the control mechanisms within the bacterium E. coli). One can then use the various measures that have been developed for characterizing networks to analyze control processes. Yet, when they zoom in to specific regions of the network, they often return to a representation of control mechanisms as at a higher level than those they control. The upshot is that, as with mechanistic levels, understanding control in terms of levels is useful locally but is difficult to implement globally.

In this section I have shown that control mechanisms can be productively viewed as organized in terms of levels, with high-level controllers operating on those at lower levels. I have emphasized that in biology control often deviates from the hierarchical model. As it does so, the simple ordering of control processes begins to break down. Overall, there does not seem to be a simple hierarchical organization of control mechanisms in biology, but generally a complex heterarchy. Nonetheless, locally researchers can usefully represent control mechanisms in local contexts in terms of levels.

4 Conclusion: Two Restricted Notions of Levels in Biology

Talk of levels is common in biology and philosophical discussions of biology. But the term *level* is used in many ways. I have focused on two conceptions of levels often invoked in biology, one associated with mechanistic explanation and one with control. Each plays an important role in the explanatory endeavors of biologists. But neither conception allows one to identify a stratified set of levels extending across biological phenomena.

Mechanistic explanations are inherently interlevel—an investigator explains a phenomenon mechanistically by decomposing the mechanism taken to be responsible for it into its components and showing how, when appropriately organized, they together generate the phenomena. In the modification of the standard account of mechanistic explanation that I have advanced, mechanistic explanations decompose the mechanism into components that together constrain the flow of free energy to produce the phenomenon being explained. On either conception of mechanisms, components belong to a lower level than the mechanism. Components can, in turn, be decomposed into other components at a yet lower level. This conception of levels is sufficient to capture what biologists have in mind in referring to bottom-up and top-down explanation without encountering problems of over-determination since the relation between mechanistic levels is not causal. The effects of causal forces impacting on a mechanism will appear at each level of decomposition—they will show up in the mechanism and in at least one of the parts of the mechanism. Mechanistic explanations appeal to the activities of component entities within the mechanism to explain what the mechanism does, but these do not cause the operation of the mechanism. Rather, the activity of the mechanism as a whole just is the organized activity of its components. But, as I have emphasized, different ways of carrying out the decomposition will result in identifying different levels. The explanatory project of appealing to one level to explain what is characterized at another is dependent on adopting a particular decomposition. Decomposition-dependent mechanistic levels are sufficient for the purposes of mechanistic explanation in biology; however, they do not give rise to a general stratification of levels in biology. One consequence is that it mechanistic levels do not allow for identifying an objective bottom level in terms of which all higher-level phenomena can be explained.

A concern with control also looms large in biology and introduces a different conception of levels. Control mechanisms that act on flexible constraints in other mechanisms are commonly represented as at a higher level than those other mechanisms. The relation between control mechanisms and the mechanisms they control is not compositional but causal. Since there is not a compositional relation between control mechanisms and those they control, a causal relation between levels does not result in over-determination—the control mechanism produces its own effect, distinct from other causes, on the mechanism it controls. This alters what action the mechanism performs on inputs. A hierarchical relation can be identified between a control mechanism and the mechanism it controls. But as researchers identify other control relations operating in the same system, the local hierarchy gives rise to a larger-scale heterarchy as multiple controllers act on the same controlled mechanism, and higher-level controllers procure information from lower-level components. As a result, one cannot stratify control mechanisms into a single layering of levels. Often researchers give up on a hierarchical representation and instead develop network representations identifying the multiplicity of ways control mechanisms impinge on each other.

Neither biological inquiries into mechanisms nor into control give rise to a global differentiation of levels in biology. While this may seem unsatisfactory to philosophers seeking a unified perspective on the natural world that includes biological phenomena, it doesn't impair biological practice. Such practice typically starts with a specific phenomenon. Whether one's goal is to explain the phenomenon in terms of the parts of a mechanism or in terms of how the mechanism is controlled within the organism, researchers do not need to appeal to a general stratification of levels. It suffices that locally they can make sense of the constitution of the mechanism or the set of control mechanisms operating on it. As one includes more—either by going to yet lower levels, or identifying multiple control mechanisms—the relations get messy, and it is difficult to maintain a coherent conception of levels. But such a globally coherent conception is not needed by biologists. If for other purposes an investigator seeks to identify stratified levels extending across biological phenomena, they will need to appeal to a different source than mechanistic explanations or control processes.

References

Bechtel, W. (2022). Reductionistic explanations of cognitive information processing: Bottoming out in neurochemistry. *Frontiers in Integrative Neuroscience*, *16*, 944303.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*(2), 421–441.

Bechtel, W., & Bollhagen, A. (2021). Active biological mechanisms: Transforming energy into motion in molecular motors. *Synthese*, *199* (5–6): 12705–12729.

Bechtel, W., & Richardson, R. C. (1993/2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Cambridge, MA: MIT Press. 1993 edition published by Princeton University Press.

Betts, J. G., Young, K. A., Wise, J. A., Johnson, E., Poe, B., Kruse, D. H., Korol, O., Johnson, J. E., Womble, M., & DeSaix, P. (2013). *Anatomy and Physiology*. OpenStax.

Bich, L., & Bechtel, W. (2022). Control mechanisms: Explaining the integration and versatility of biological organisms. *Adaptive Behavior*, *30*(5), 389–407. doi:10.1177/10597123221074429

Bjursten, L. M., Norrsell, K., & Norrsell, U. (1976). Behavioural repertory of cats without cerebral cortex from infancy. *Experimental Brain Research*, *25*(2), 115–130. doi:10.1007/BF00234897

Churchland, P. S., & Sejnowski, T. J. (1988). Perspectives on cognitive neuroscience. *Science, 242*, 741–745.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. New York: Oxford University Press.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy*, *22*, 547–563.

Finer, J. T., Simmons, R. M., & Spudich, J. A. (1994). Single myosin molecule mechanics: piconewton forces and nanometre steps. *Nature*, *368*(6467), 113–119. doi:10.1038/368113a0

Glennan, S. (2017). *The new mechanical philosophy*. Oxford: Oxford University Press.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & Group, T. C. S. R. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Science*, *19*(1), 46–54. doi:10.1016/j.tics.2014.10.004

Hooker, C. A. (2013). On the import of constraints in complex dynamical systems. *Foundations of Science*, *18*(4), 757–780. doi:10.1007/s10699-012-9304-9

Huxley, A. F., & Niedergerke, R. (1954). Structural changes in muscle during contraction—Interference microscopy of living muscle fibres. *Nature*, *173*(4412), 971–973.

Huxley, H. E. (1969). The mechanism of muscular contraction. *Science*, *164*(3886), 1356–1365.

Huxley, H. E., & Hanson, J. (1954). Changes in the cross-striations of muscle during contraction and stretch and their structural interpretation. *Nature*, *173*(4412), 973–976.

Katz, P. S. (1999). What are we talking about? Modes of neuronal communication. In P. S. Katz (Ed.), *Beyond neurotransmission: Neuromodulation and its importance for information processing* (pp. 1–28). New York: Oxford.

Kauffman, S. A. (2000). *Investigations*. Oxford: Oxford University Press.

Kim, J. (1998). *Mind in a physical world*. Cambridge, MA: MIT Press.

Lanner, J., Georgiou, D. K., Joshi, A., & Hamilton, S. (2010). Ryanodine receptors: structure, expression, molecular details, and function in calcium release. *Cold Spring Harbor Perspectives in Biology, 2*(11), a003996.

Leng, G. (2018). *The heart of the brain: The hypothalamus and its hormones*. Cambridge: MIT Press.

Lymn, R. W., & Taylor, E. W. (1971). Mechanism of adenosine triphosphate hydrolysis by actomyosin. *Biochemistry*, *10*(25), 4617–4624. doi:10.1021/bi00801a004

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1–25. doi:10.1086/392759

McCulloch, W. S. (1945). A heterarchy of values determined by the topology of nervous nets. *The Bulletin of Mathematical Biophysics*, *7*(2), 89–93. doi:10.1007/BF02478457

Pattee, H. H. (1991). Measurement-control heterarchical networks in living systems. *International Journal of General Systems*, *18*(3), 213–221.

Rayment, I., Holden, H. M., Whittaker, M., Yohn, C. B., Lorenz, M., Holmes, K. C., & Milligan, R. A. (1993). Structure of the actin-myosin complex and its implications for muscle contraction. *Science*, *261*(5117), 58–65.

Rayment, I., Smith, C., & Yount, R. G. (1996). The active site of myosin. *Annu Rev Physiol, 58*, 671–702. doi:10.1146/annurev.ph.58.030196.003323

Redgrave, P., Vautrelle, N., & Reynolds, J. N. (2011). Functional properties of the basal ganglia's re-entrant loop architecture: selection and reinforcement. *Neuroscience, 198*, 138–151. doi:10.1016/j.neuroscience.2011.07.060

Shik, M. L., & Orlovsky, G. N. (1976). Neurophysiology of locomotor automatism. *Physiological Reviews*, *56*(3), 465–501. doi:10.1152/physrev.1976.56.3.465

Sklar, L. (2013). *Philosophy and the foundations of dynamics*. Cambridge: Cambridge University Press.

Winning, J., & Bechtel, W. (2018). Rethinking causality in neural mechanisms: Constraints and control. *Minds and Machines*, *28*(2), 287–310. doi:10.1007/s11023-018-9458-5

8
From Analogies to Levels of Abstraction in Cognitive Neuroscience

Mazviita Chirimuuta

1 Levels of Organisation and Levels of Abstraction

The term *levels of explanation* can be taken either as an epistemic notion, a metaphysical one, or both (as the various chapters in this volume demonstrate). Some influential characterisations of levels in neuroscience have centred *levels of organization*, which is a metaphysical notion, the idea being that the brain and nervous system by themselves, independently of scientific representation, divide into hierarchically arranged structures. For example, Churchland and Sejnowski (1988) presented a widely reproduced diagram of levels of scale, in which the various working parts of the central nervous system (CNS), from molecules and synapses to brain maps and systems, are each associated with a characteristic scale, from 1 angstrom up to 1 metre for the CNS as a whole. A widely known proposal for levels of organisation comes from Craver (2007: Ch. 5), who depicts the central nervous system as a hierarchy of stacked mechanisms, such that a cognitive phenomenon like memory consolidation is conceived as a system of interacting parts, and where the parts of this mechanism are themselves sub-mechanisms, and so forth, bearing part/whole relationships across levels.

One other proposal for levels of organisation has a strong connection with the theme of this chapter, since it too proposes a distinct computational level grounded in an analogy with man-made computers. This is the so-called received view (Piccinini and Craver 2011), popularized by Jerry Fodor and others from the 1960s onwards (e.g., Fodor 1997). The view, essentially a functionalist ontology of mind, has it that the mind is its own level, distinct from neurophysiology, and that the brain is the realizer of the mind, just as the hardware of a computer is the realizer of its software. A feature of the view is that psychology may proceed autonomously from neurophysiological investigation, and that psychology has its own characteristic explanatory form, proceeding via *functional analysis*. Given that these ontological levels are in part characterised in terms of modes of explanation, we have an indication that the separation between epistemic and metaphysical notions is less clear cut than it first appears. This will be a recurrent theme of the chapter.

Mazviita Chirimuuta, *From Analogies to Levels of Abstraction in Cognitive Neuroscience* In: *Levels of Explanation*.
Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Mazviita Chirimuuta 2024.
DOI: 10.1093/oso/9780192862945.003.0009

Moving onto the epistemic side, the core idea is that different forms of explanation, which for reasons that will become apparent below I characterize as different *levels of abstraction*, are applicable to what is ontologically one and the same target, e.g., the retina. As such, the separation into levels is not pre-given in nature, but is a product of scientific research and representation. Under this heading we find Marr's 'levels of explanation',[1] and also the proposals of cognitive scientists Zenon Pylyshyn and Allen Newell (Elber-Dorozko and Shagrir 2019), not to mention the three-stance system of Daniel Dennett (1987).

As is well known, Marr's framework is introduced in the first chapter of the neuroscientist's posthumously published book, *Vision*. The three levels, given on p. 25 are:

(1) Computational theory
(2) Representation and algorithm
(3) Hardware implementation

The 'top level' computational theory gives an abstract characterisation of the performance of a system in terms of its generating a mapping of an input to an output. In addition, characterisation at this level shows how that performance is related to environmental constraints and behavioural goals. Thus, the first level is to provide a functional characterisation in both senses of the word: explicating a mathematical input-output mapping, and also illuminating the utility of the performance.[2] The middle level involves specification of the format for representation of the inputs and outputs, and of the algorithm that transforms one into the other. The bottom level describes how the representations and algorithm are physically realised, for example in the electronic components of a computer vision system, or in the neurons of an animal's retina.

In Section 2 I will say more about how analogies with machines motivate this three-level system, and why they are essential in the interpretation of it. Here we should note that Marr's proposal carries on from a discussion of the limitations of reductionist approaches to explaining the visual system—attempts to understand how neural activity gives rise to useful perceptions of the environment by way of careful study of the anatomy and physiology of neurons. In effect, the reductionist is restricted to the bottom level of explanation. Marr (1982: 27) describes this approach as equivalent in futility with the attempt to understand bird

[1] The view of Maley (2021) is that Marr's levels are *not* levels of abstraction. This seems to be based on the point that there can be degrees of abstraction within a Marrian level (intra-level). I do not think this rules out my point that the levels also differ in degree of abstraction when compared against each other (inter-level).

[2] Egan's (2014) interpretation focuses just on the mathematical sense, but as Shagrir and Bechtel (2017) point out, the ecological interpretation of the visual system's performance seems to be just as central to the Computational theory. Shagrir and Bechtel put levels (2) and (3) together because internally focused. But on other hand, (1) and (2) go together because both abstract away from neural details.

flight just through the examination of feathers. As he asserts in the preamble to the three levels, '[a]lmost never can a complex system of any kind be understood as a simple extrapolation from the properties of its elementary components' (Marr 1982: 19). The basic complaint against reductionism is that this is a strategy that quickly gets the investigator overwhelmed with details whose significance cannot be assessed because she lacks knowledge of the overall functionality of the system, and therefore has no working hypothesis about how the elementary components contribute to global properties and behaviour. The shape of the forest is invisible because there are so very many leaves. The introduction of the two additional levels of explanation allows for lines of investigation that prioritise general questions about the system's functionality and operations independently of investigation into implementational details. The upper two levels are analyses that abstract away from the complications inherent to the material system. Ideally, the results of these upper-level investigations provide a map of what to look for in the concrete system, and a guide to interpreting the material details, even though the levels are only 'loosely related' (Marr 1982: 25).

One of the virtues of Marr's framework, highlighted by later researchers, is that it offers this strategy for simplification.[3] For example, Dana Ballard (2015: 13) writes that it 'opened up thinking about the brain's computation in abstract algorithmic terms while postponing the reconciliation with biological structures'. Speaking of level schemas more generally, Ballard emphasizes that '[b]y telescoping through different levels, we can parcellate the brain's enormous complexity into manageable levels' (2015: 18). But, of course, Marr was not the first theorist of living systems to have had the idea that explanation should not be restricted to the reductionist analysis of material hardware, and that function-first approaches sometimes need to be taken. In Aristotle's natural philosophy, questions of overall order ('form') and function ('finality') are primary, and questions about the material composition of organisms come second. Aristotle's framework was highly metaphysical, with forms and final causes as ontological posits, and was castigated as anthropomorphic by proponents of modern, mechanical natural philosophy.[4] Yet, arguably there is an Aristotle-shaped hole in the heart of mechanistic biological science precisely because it is rarely possible to show how complex phenomena in living systems are produced by rearrangements of 'elementary components'.

[3] Of course, the details of Marr's framework have been criticised by later researchers, such as Love (2021), who argue for a greater number of levels. Gurney (2009) proposes a four-level framework which is incidentally more similar to one proposed by Marr in a 1976 technical report.

[4] Mechanical natural philosophy is heterogenous and can be hard to characterise. But in contrast with the Aristotelian-scholastic natural philosophy, it is certainly more reductionistic (Pasnau 2011: 50). It is helpful to consider Leibniz's characterisation of the 'mechanical':

> everything must happen in the bodies in such a way that it is possible to explain it distinctly from the very nature of the bodies, that is, from the size, the figure and the laws of motion: this is what I call 'mechanical'. (From Leibniz's Animadversiones, quoted in Nunziante 2020: 15)

We would now call this a 'bottom up' mode of explanation.

Rather than reinstate Aristotelian metaphysics, the strategy has been to invoke comparisons and analogies with engineered systems for which the notions of function and design are germane.[5] By use of artefact analogies in explanations of living systems, the invocation of function and finality can be confined to the epistemic side, bypassing scepticism about the existence of evolved teleo-functions, and metaphysical worries about whether higher levels of organisation can be said to have some kind of priority over their component parts. A very deflationary approach to functional considerations is put forwards by Craver (2013), in which the functional 'perspective' is needed to help delineate a mechanism, setting it apart from the countless interrelated entities and processes that make up a living system. But at the end of the day the ascription of function depends on a human-dependent perspective, in contrast with the fully mind-independent status of the causal nexus of mechanisms when they are considered without functional descriptions.[6]

2 The Artifact Analogies

It is not appropriate to attribute the very deflationary stance, noted at the end of the previous section, to Marr himself. For one thing, he is not a philosopher and does not present his work in terms that clearly support an interpretation on this point. What we can say is that the general impression given by Marr's presentation is that he does not care to set a division between engineered and living systems, between those that have (computational) functions, properly speaking, and those for which it is only a heuristic posit.[7] Furthermore, we will see by the end of this section that in the actual deployment of the framework within computational neuroscience, researchers have not managed to avoid 'metaphysical creep'—there tends to be some at least tacit commitment to a hierarchy of levels of organisation within the brain, existing independently of scientific description.

[5] This is an observation made by neurophysiologist, Jerome Lettvin:

> Ever since biology became a science at the hands of biochemists it has carefully avoided or renounced the concept of purpose as having any role in the systems observed … Only the observer may have purpose, but nothing observed is to be explained by it. This materialist article of faith has forced any study of process out of science and into the hands of engineers to whom purpose and process are the fundamental concepts in designing and understanding and optimizing machines.' (Lettvin, interviewed in Anderson and Rosenfeld 1998: 13)

See essays in Allen, Bekoff, and Lauder (1998) for further discussion of 'purpose' in modern biology.

[6] Kant (1790/1952) is an early proponent of the epistemic approach to form, function, and finality. In the third *Critique*, the causal-mechanical characterization of a living body is 'constitutive', whereas the teleological one is merely 'regulative'.

[7] I take a different tack here. In my recommendations for how best to interpret computation-level descriptions in neuroscience, I argue that the boundary does need to be drawn between these two kinds of things, and that computational descriptions are not literally true of neural systems (Chirimuuta 2021, 2024).

A striking feature of Marr's presentation is that in the first instance it relies exclusively on examples of information processing machines. Cases from within neuroscience are mentioned only after a complete account of the three levels has been given, without there being any comment on this transition. The primary illustration of the levels comes by way of a cash register, an adding machine. At the computational level, the task is to find out '*what* the device does and *why*' (Marr 1982: 22).[8] This means specification of the arithmetical theory of addition, as well as an account of the functional role of the machine for adding up charges in a shop. We learn that the second level characterisation involves showing how numbers are represented in the device (e.g., Arabic or Roman notation), and specifying the algorithm used to work out the total bill. The implementation level involves characterisation of the 'physical substrate' which runs the algorithm. A point Marr (1982: 24) emphasises is that the same algorithm can be realized in very different materials. This also goes for the relationship between the top two levels: one and the same computational task can be achieved by a range of different algorithms. This is why the levels are only 'loosely related' (p. 25)—a discovery at one level cannot reliably pre-specify what will be found at the level below.

We might speculate that Marr leans on artefacts for purposes of exposition just because the core concept of each of these levels comes out especially clearly in cases like the cash register. But then we ought to wonder why it is that it is harder to get a grip on how to define these levels in neuroscience, even though the framework is intended for use there. We can discern a deeper reason for the primacy of machines in Marr's exposition if we consider Dennett's observation that the three levels actually schematise the stages taken in the engineering of a complex information processing system. Dennett (1995: 682) writes,

> Marr's obiter dicta [passing words] on methodology gave compact and influential expression to what were already reigning assumptions in Artificial Intelligence. If AI is considered as primarily an engineering discipline, whose goal is to create intelligent robots or thinking machines, then it is quite obvious that standard

[8] To reinforce this point about the primacy of artifacts, note that Marr does not use the neutral language of 'things' or 'systems' but refers specifically to a 'device' here. We find this also in the legend for the summary table: 'The three levels at which any *machine* carrying out an information-processing task must be understood' (p. 25 emphasis added). Cf. 'the different levels at which an information processing *device* must be understood before one can be said to have understood it completely' (p. 24 emphasis added).

Later in the book, when again summarising the three levels as applied to the visual system, it is interesting that the terms 'machine' and 'machinery' are still used:

> The human system is a working example of a machine that can make such descriptions, and as we have seen, one of our aims is to understand it thoroughly, at all levels: What kind of information does the human visual system represent, what kind of computations does it perform to obtain this information, and why? How does it represent this information, and how are the computations performed and with what algorithms? Once these questions have been answered, we can finally ask, How are these specific representations and algorithms implemented in neural machinery? (Marr 1982: 99)

> engineering principles should guide the research activity: first you try to describe, as generally as possible, the capacities or competences you want to design, and then you try to specify, at an abstract level, how you would implement these capacities, and then, with these design parameters tentatively or defeasibly fixed, you proceed to the nitty-gritty of physical realization.

The point here is that the three levels of explanation are an expression of three broad steps in the *forward engineering* of a machine with some functionality equivalent to a cognitive capacity in an animal. It is then not surprising that the different levels are more easy to illustrate with an example of *reverse engineering* some such device.

The issue I am highlighting here is that artefacts are the foundational cases for Marr's framework, and the application to neuroscience occurs via an analogical transfer to brains, systems which are putatively similar to computing ones. Researchers habitually think of brains, just like the artefacts, as taking in inputs (e.g., from sensory organs), implementing some algorithms, and sending an output (e.g., a motor command).[9] The importance of this analogy comes out in Dennett's characterisation of what his approach has in common with that of Newell and Marr, namely:

> (1) Stress on being able (in principle) to specify the function computed (the knowledge level or intentional level) independently of the other levels.
> (2) An optimistic assumption of a specific sort of functionalism: one that presupposes that the concept of the function of a particular cognitive system or sub-system can be specified. (It is the function which is to be optimally implemented.)
> (3) A willingness to view psychology or cognitive science as reverse engineering in a rather straightforward way. Reverse engineering is just what the term implies: the interpretation of an already existing artifact by an analysis of the design considerations that must have governed its creation. (Dennett 1995: 683)

Dennett's articulation of the reverse engineering methodology, his *design stance*, comes with strict assumptions of optimality and adaptationism in evolved systems that we need not attribute to the scientific practice. In my view, the essential point about the reverse engineering methodology is that it treats the biological object by

[9] E.g., Marcus and Freeman (2015: xiii):

> The brain is not a laptop, but presumably it is an information processor of some kind, taking in inputs from the world and transforming them into models of the world and instructions to the motor systems that control our bodies and our voices.

See Chirimuuta (2021, 2024) on why this practice should be interpreted as resting on a loose analogy rather than strict functional similarity between computer and brain.

analogy with a man-made thing, and in this way attempts to make it intelligible by showing how it operates according to principles that make sense from the perspective of a person designing things; in other words, by treating it as if it were an artefact, the scientist can explain it in terms of the practical rationality of causal means being used to produce useful effects.

We should appreciate that there are two levels of analogy, so to speak. Superficially, the analogy just holds between certain organs of living bodies and man-made devices that have a rough functional equivalence with them—the brain and a computer, the heart and a pump. But the deeper and more general point is that there is an analogy being invoked between the systematic organisation of parts and processes through which organs generate their functional effects, and the parts and processes set in place by a human engineer in order for a device to achieve the desired effect. An artefact is intelligible to the extent that its operations are the manifestations of the instrumental rationality through which its human makers achieve their ends.[10] A similar kind of intelligibility is tacitly assumed for the biological object. This becomes clearer when we consider functional analysis, which is a general schema for reverse engineering.

When a reverse engineer is presented with a machine or biological organ, their task is, in general terms, to show how the capacity (or functional disposition)[11] of the entire thing can be redescribed in terms of some simpler capacities (or dispositions). These simpler capacities may already be reverse engineered, and therefore intelligible, in which case no further analysis is needed, or they may themselves need to undergo a redescription in terms of simpler capacities, and so on, down through different levels of analysis. This is how Cummins (2000: 125) puts it:

> Functional analysis consists in analyzing a disposition into a number of less problematic dispositions such that programmed manifestation of these analyzing dispositions amounts to a manifestation of the analyzed disposition. By 'programmed' here, I simply mean organized in a way that could be specified in a program or flowchart.

[10] Incidentally, this is how Kant proposes to accommodate the notion of end-directed causality within modern mechanistic science, as founded on an analogy that humans derive from their own activities in producing technical objects, guided by means-end reasoning:

> we picture to ourselves the possibility of the object on the analogy of a causality of this kind—a causality such as we experience in ourselves—and so regard nature as possessed of a capacity of its own for acting *technically*. (Kant 1790/1952: Part II, 5 / §361)

See Breitenbach (2014) and Illetterati (2014) for discussion of Kant's ideas about the analogy. A theological assumption in the background of finalism is that living beings are intelligible to the extent that they are the work of a rational Creator. Kant's proposal retains the connection between intelligibility and instrumentally rational creation, but drops the theological assumption by saying that when functions are attributed to the objects of biology, we view them *as if* they were the works of nature acting rationally.

[11] The distinction between capacities and dispositions is not relevant to my discussion here.

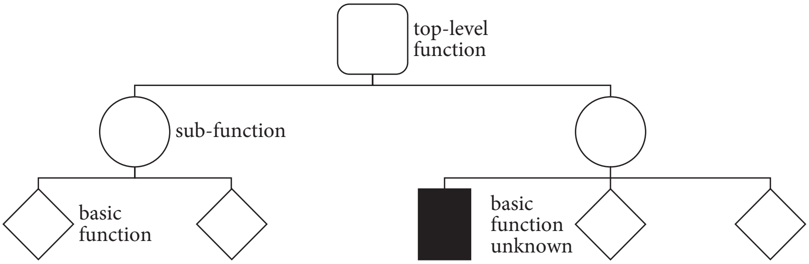

Figure 8.1 Illustrates the hierarchical, modular organisation that enables functional analysis. The top-level function can be decomposed into two sub-functions, which are themselves subject to decomposition into basic functions. Since the number of interactions between functional components is small, their inner workings can be 'black-boxed' (deliberately or by necessity ignored) so that higher-level functional explanation can proceed independently of knowledge of the lower-level details.

Cummins tells us that functional analysis is best illustrated through the example of assembly line production. The capacity of the whole factory to produce a certain product is broken down into a sequence of simpler sub-tasks which are intelligible without further analysis. When biologists and psychologists offer functional explanations of the capacities of the body and mind, Cummins (1975: 760–761) argues that they are following this pattern of analysis. A point that is not emphasized in Cummins's discussion is that deployment of this explanatory strategy beyond the domain of artefacts and manufacture presumes that the same kinds of intelligible forms of organisation are to be found in nature. For example, the assumption that a capacity can be analysed down into sub-capacities assumes that the whole system is a composite of encapsulated, specialized modules whose workings can be understood independently of their context within the whole.[12]

The link between this reverse engineering methodology in cognitive science and neuroscience and simplification of the brain becomes apparent if we focus on the importance of encapsulation in functional analysis (see Figure 8.1). When a system is described in this way, the payoff is that at any given level of analysis the component modules can be treated as black boxes whose inner workings are either unknown or ignored, since the only information relevant to the current level of analysis is the input-output profiles of the modules. Descent to a lower level of analysis involves opening the black boxes and seeing how their inner workings can be accounted for in terms of the functional capacities of their components. But

[12] As Schierwagen (2012: 144) points out, the concept of a module itself originates from engineering: it denotes the process of decomposing a product into building blocks, modules, with specified interfaces, driven by the designer's interests and intended functions of the product.

for many explanatory purposes, lower-level details can safely be kept out of view, which is why this methodology offers a handy simplification.

To illustrate this point, I will make use of an example from computing given by Ballard (2015: 14ff.). Most people who program computers only ever use a high-level programming language such as Python. But the terms of this high-level language are actually black boxes which unpack into more complicated expressions in a lower-level assembly language. These lower-level terms themselves unpack into instructions in machine code, which are not mere shorthand for the higher-level commands. For a programme to be carried out, it needs to be translated down into lower-level languages, 'closer to machine's architecture'. But this is all done behind the scenes, and the ordinary coder can comfortably stick with description of the computation in the compact, highest-level language. The point of Ballard's example here is to argue that there is a tight analogy between the computer and the brain, which he thinks can be described similarly in terms of 'levels of computational abstraction'.

A concern that arises at this point is that the levels framework equivocates between an epistemic and a metaphysical proposal, between *description* in terms of levels of abstraction and the *positing* of levels of organisation. For a computer is not merely a system that can be described at these different levels; the levelled hierarchy really is a feature of its hardware construction and software design. And so we find a metaphysical creep in Ballard's application of this idea of levels of abstraction to neuroscience:

> we are unlikely to get away with a 'flat' neural computation *description*. The far more likely arrangement is that *the brain is composed* of many more abstract neural networks that leverage the results of less abstract networks in the process of getting things done. (Ballard 2015: 21 emphasis added)

In this passage Ballard switches from talking of the need for multilevel description to an assertion of multiple levels of abstraction in the composition of the brain. How can neural networks be more or less abstract, since they are concrete cellular structures? I think what Ballard means to say is that the representations *constructed within these neural networks* are more or less abstract. Crucially, the abstraction hierarchy is posited to be there in the brain's own representations of the extra-cranial world, not just in those imposed upon it by a scientist. The proposal is that the brain is a system that, at the top level of control, ignores its own complexity, like a digital computer where the execution of a piece of code is indifferent to micro-physical fluctuations in the electronic hardware. Just as the programmer, the controller of a computer, can govern the performances of the machine while ignoring and remaining ignorant of its low-level languages and physical workings, it is supposed that the brain systems ultimately responsible for behaviour employ an abstract, high-level system of representation that is invariant to changes in the complex, low-level workings of the brain and rest of the body. If this assumption holds, there are good prospects for a relatively simple computational theory that explains how the brain governs behaviour, by way of these high-level representations.

But why would neuroscientists think that this assumption does hold, that the analogy between computer and brain is tight enough? The intelligible organization of systems as hierarchically arranged, encapsulated modules, or levels of more or less abstract representations, can be found in artefacts designed by humans, but its existence in the natural world should not be taken for granted. As far as I can determine, the foundational argument in support of this assumption comes from another analogy put forwards by Herbert Simon in the 'Architecture of Complexity' (Simon 1969, 1962).[13] In a tale of two watchmakers, Simon describes how the production of a complex system (a watch) is much more likely to be successful if the production process occurs in stages, where sub-processes in the production result in stable sub-components of the system that are assembled together at a later stage. Simon then draws an analogy between human manufacture and the evolution of complex life forms. His point is that the likelihood of evolution producing organisms of any complexity is vanishingly small unless it is the case that it comes about via the evolution of intermediate, self-standing forms that become the components of more complex organisms. Hence, he argues, it must be the case that evolved as well as manufactured complex systems are composed, hierarchically, of relatively independent sub-systems. In these *near decomposable* complex systems, there is only a weak frequency and strength of interaction horizontally between the sub-systems at any one level, and vertically across the levels of organization. This means that the sub-systems—the modular components—can usefully be studied in isolation from the rest of the system, and that the system can be studied at higher levels of organization (which we can here equate to larger scales) without attention to most of the lower-level (i.e., small-scale) details. The optimistic upshot is that evolved complex systems are scientifically intelligible through decomposition into

[13] It is interesting that Marr (1982: 102) also makes the connection between evolvability, intelligibility, and modular organisation:

> This observation [of isolated visual processing] ... is fundamental to our approach, for it enables us to begin separating the visual process into pieces that can be understood individually. Computer scientists call the separate pieces of a process its *modules*, and the idea that a large computation can be split up and implemented as a collection of parts that are as nearly independent of one another as the overall task allows, is so important that I was moved to elevate it to a principle, the *principle of modular design*. This principle is important because if a process is not designed in this way, a small change in one place has consequences in many other places. As a result, the process as a whole is extremely difficult to debug or to improve, whether by a human designer or in the course of natural evolution, because a small change to improve one part has to be accompanied by many simultaneous, compensatory changes elsewhere. The principle of modular design does not forbid weak interactions between different modules in a task, but it does insist that the overall organization must, to a first approximation, be modular.

levels and components, and that this is an alternative to intractable reductionist methodologies.[14]

Simon gives the following compact summary of the position, which shows how the claim for there being levels of organisation in nature is used to justify the methodological strategy of investigating complex systems at different levels of explanation:

> Scientific knowledge is organized in levels, not because reduction in principle is impossible, but because nature is organized in levels, and the pattern at each level is most clearly discerned by abstracting from the detail of the levels far below … And nature is organized in levels because hierarchic structures … provide the most viable form for any system of even moderate complexity. (Simon 1973: 26–27; quoted in Wu 2013: 282)

In the next section I will give some critical commentary on this approach. Before moving on it is worth saying more about the role of all these ideas as a simplifying strategy in neuroscience.

Reductionist methodologies can be successful for relatively simple systems. The task of the research is to acquire sufficient information about the elementary components, and their interaction, to yield an explanation of the behaviour of the whole system. This is a 'flat', as opposed to multilevel, approach. Once there is enough complexity that the amount of information about elementary components and the interactions that can feasibly be dealt with (in models or theory) is much less than what is required for explanation of the system's behaviour, then a multilevel approach is needed. The common virtue of all of the multilevel approaches discussed above—from Marr, Ballard, and Simon—is that they offer a guide for how to abstract away from low-level details and how to set about work on top-down explanations when bottom-up, reductionist approaches are intractable, even if possible in principle. These three scientists are all proponents of computational explanations of how the brain gives rise to cognition, and this kind of explanatory practice is favoured because, they argue, it does not require that much attention be paid to the details of neurophysiology which would otherwise threaten an overwhelming complexity.

An additional feature of computational explanations is that they assert an equivalence between organic and artificial systems, so long as they are computing the same functions. The idea here is familiar to philosophers under the heading of multiple realisation. A mechanical cash register, an electronic calculator, and a human brain region can all be said to be doing the same computation when

[14] See Bechtel and Richardson (2010) for further discussion of methods for investigating near decomposable systems.

adding up a particular sum, even though the physical substrates are so different. The benefit of this for neuroscientific research is that it justifies the substitution of actual neural tissue with *relatively* simple computational models, such as artificial neural networks (ANNs), as objects of investigation.[15] A goal of various neurocomputational research projects has been to create models of brain areas *in silico* that will yield confirmatory or disconfirmatory evidence for theories of cognition and pathology, where traditional experimental approaches are untenable because it is not possible to make the required interventions on actual neurons.[16] Even though large ANNs are themselves rather complicated and hard to interpret, they are at least more accessible to (simulated) experimental interventions, such as lesioning of individual nodes.

Aside from the specifics of computational explanation (explanation via analogy between brains and computers), one of the general implications of the artefact analogy is that the nervous system is composed of relatively encapsulated working parts (modules) or functional components. This also supports the 'black-boxing' of neural details. As Haugeland (1978: 221) relates,

> if neurons are to be functional components in a physiological system, then some specific few of their countless physical, chemical, and biological interactions must encapsulate all that is relevant to understanding whatever ability of that system is being explained.

One way to think about the importance of *neuron doctrines* in the history of the discipline—theories that posit individual neurons as the basic anatomical and functional units of the nervous system—is that they facilitate this simplifying strategy, even while departing from many of the observable results on the significance of sub-neuronal and non-neuronal structures and interactions.[17] Moreover, we should note also that this black-boxing can be employed to achieve abstract representations of functional components other than individual neurons. For example, Hawkins, Ahmad, and Cui (2017) present a model of cortical columns (an area of the cortex approximately 1 mm^2, c. 100 neurons) in which they are treated as stereotyped input-output units which are fed sensory and location information, and can learn to recognise objects when joined together in small networks.

[15] Amongst very many studies, a good example is Mante et al. (2013), discussed in Chirimuuta (2018).

[16] The Blue Brain Project is the most notorious instance of the ambition to create a large scale *in silico* brain replica. For commentary see Koch and Buice (2015) and Mahfoud (2021).

[17] See Bullock et al. (2005) and Cao (2014) on the empirical inadequacy of the neuron doctrine. Barlow (1972) is a great example of its role in explanatory simplification. Larkum (2022) explains how such views mis-state the role of dendrites in brain function.

3 Limitations of the Analogies

I have argued that the dominant multilevel approaches in neuroscience rest on the assertion of there being a close similarity between the multilevel organisation of artefacts such as computers, and the brain, an evolved organ whose organisational 'plan' is far less well characterized than that of the machine, and remains a matter of controversy. This prompts consideration of the difficulties that the multilevel approach faces, to the extent that the claim for similarity can be challenged. If the comparison between brain and computer is at best a loose analogy, in which the dissimilarities between the two are of equal importance or even outnumber the similarities, then the levelled approach might sometimes be a hindrance in the project of explaining how brain activity gives rise to cognition.

The first concern to bring up here is that the case for encapsulation in the nervous system is fairly weak. This was pointed out decades ago by Haugeland, in the passage following on from the one quoted above:

> [encapsulation] is not at all guaranteed by the fact that cell membranes provide an anatomically conspicuous gerrymandering of the brain. More important, however, even if neurons were components in some system, that still would not guarantee the possibility of 'building back up.' Not every contiguous collection of components constitutes a single component in a higher-level system; consolidation into a single higher component requires a further encapsulation of what's relevant into a few specific abilities and interactions—usually different in kind from those of any of the smaller components. Thus the tuner, pre-amp and power amp of a radio have very narrowly specified abilities and interactions, compared to those of some arbitrary connected collection of resistors, capacitors, and transistors. The bare existence of functionally organized neurons would not guarantee that such higher-level consolidations were possible. Moreover, this failure of a guarantee would occur again and again at every level on every dimension. There is no way to know whether these explanatory consolidations from below are possible, without already knowing whether the corresponding systematic explanations and reductions from above are possible—which is the original circularity. (Haugeland 1978: 221)

It is interesting that Haugeland focuses on the possibility of a strong disanalogy between the organization of the nervous system and that of a human-designed artefact, a radio. Whereas it is a feature of the design of a radio that higher-level sub-components (the tuner, pre-amp, and power amp) are made up of careful arrangements of lower-level sub-components (resistors, capacitors, and transistors), and themselves have narrowly specified capacities and input-output profiles, it should not be assumed that collections of neurons consolidate into higher-level sub-components in this way, and that explanations of the neural basis of cognition can safely be restricted to the higher levels. I will now discuss two reasons to be sceptical that the analogy holds. The first relates to the potential importance of low-level activity, the second brings up the difference between hierarchical, designed systems and evolved ones.

It is an open possibility that cognition is the product of dense interactions across a number of levels or scales and is not restricted to a high level of computational abstraction, as Ballard would have it. The cognitive properties of the brain may be enmeshed in its material details, in a way not congenial to Marr's vision of there being computational and algorithmic/representation levels that are only loosely related to the implementational one.[18] A reason to give credence to these possibilities comes from consideration of the fact that biological signalling, a general feature of living cells, is the omnipresent background to neuronal functionality. The low-level details of neuronal activity can themselves be characterized as doing information processing, and are not merely the hardware implementors of the system's global computations, or bits of infrastructure keeping the system running. This is an argument put forwards by Godfrey-Smith (2016: 503):

> This coarse-grained cognitive profile is part of what a living system has, but it also has fine-grained functional properties—a host of micro-computational activities in each cell, signal-like interactions between cells, self-maintenance and control of boundaries, and so on. Those finer-grained features are not merely ways of realizing the cognitive profile of the system. They matter in ways that can independently be identified as cognitively important.

The point is that in an electronic computer there is a clean separation of the properties of the physical components that are there holding the device together, and the ones involved specifically in information processing.[19] This is how the machine has been designed, whereas in the brain this is not the case—it is not clear cut which entities within the brain, and which of their properties, are responsible for information processing, and which are the infrastructural background.[20] In

[18] Maley (2021) reaches a similar conclusion by arguing that in analogue computation algorithm/representation and implementation collapse into a single level, and that neural systems are better described as analogue rather than digital.

[19] Moreover, an interesting argument for the brain-computer *disanalogy* comes from Conrad (1989), who argued that brains are able to exploit many more of their physical interactions for information processing than is possible with a programmable computer, because for a computer to be programmable a set of instructions in programming language must map onto a limited set of physical interactions, whereas living, non-programmed systems are not bounded in this way. Conrad calls it the 'programmability/efficiency tradeoff', and it leads him to conclude that information processing in the brain occurs in an efficient, medium-*dependent* way, such that structure and function are linked and relevant interactions occur at many scales. For commentary and some criticisms, see Zeigler (2002).

[20] For example, glial cells—the very numerous kinds of brain cells that do not generate action potentials—were long thought to be providing metabolic support, but not involved in cognition. This does not appear to be the case (Cao 2014), but the challenge of integrating glia into computational theory is immense.

addition to the 'coarse-grained' computations that might be attributed on the basis of the whole animal's psychology and behaviour, Godfrey-Smith argues that there are a countless number of 'micro-computational activities' in cells, which are not unrelated to global cognition. If in the brain metabolism, cell maintenance, and global (i.e., person-level) cognitive functions are enmeshed together, then low-level material details about neural tissue, such as the specific chemical structures of the many kinds of neurotransmitter, and the thousands of proteins expressed at synapses (Grant 2018), probably do matter to the explanation of cognition. They cannot be safely discounted with the same confidence as merited in aeronautics, when air is treated as a continuous fluid and molecular details are left unrepresented.[21] It should be pointed out that Simon's notion of a near-decomposable system granted that there were interactions between levels, but took their effects to be small relative to those within levels, and therefore discountable. The question at issue here—and it is an open empirical question for neuroscience—is whether these cross-level interactions are as negligible as has been assumed.

We also saw that Simon gives an in-principle argument for the existence of hierarchical organisation in complex living systems which would, if accepted, justify the exclusion of low-level details for the purposes of most explanations of whole system behaviour. However, the strict analogy this argument supposes, between human manufacture and the processes of evolution, calls for scrutiny. Bechtel and Bich (2021) argue that hierarchical structures, with their neat pyramidal arrangement of superordinate and subordinate levels, are less likely to evolve than *heterarchical systems*, which have a more haphazard arrangement of horizontal and vertical interconnections, meaning that one component of the system is open to significant influence from components at other levels (they are not just 'loosely related'), and there is no top-level locus of control, as posited by Ballard (2015: 242) in his comparison between control in robots and humans. The reason for the hypothesized predominance of heterarchical systems is that evolution is not like a smooth, linear, process of design and manufacture, but is full of processes comparable with what those engineers would call 'tinkering' and 'kludging'.[22] A common occurrence in evolution is that a trait that is adaptive because serving one function is coopted for another, and so it is not obvious what *the* function of the trait is in the subsequent system. Cooption and functional multi-tasking are reasons why evolved systems have the heterarchical character of interactions ranging across levels. Generally speaking, to the extent that evolution is 'inelegant' and divergent from the designs that would be considered rational and perhaps optimal by a human engineer, there is an obstacle to understanding organic systems through

[21] Lillicrap and Kording (2019) also argue against the comparison between coarse-graining methods in physics and computational explanation in neuroscience.

[22] We should note here that Ballard's representation of software systems as neat and pyramidal is itself an idealisation, since large programmes like Microsoft Word are themselves the result of years of tinkering and kludging of previous versions of the code.

reverse engineering. This is a point made by Patricia Kitcher (1988) in relation to Marr's levels, and is reiterated by these biologists more recently:

> deep degeneracy at all levels is an integral part of biology, where machineries[23] are developed through evolution to cope with a multiplicity of functions, and are therefore not necessarily optimized to the problem that we choose to reverse engineer. Viewed in this way, our limitation in reverse engineering a biological system might reflect our misconception of what a design principle in biology is. There are good reasons to believe that this conclusion is generally applicable to reverse engineering in a wide range of biological systems. (Marom et al. 2009: 3)

Of course, Dennett, in his advocacy of the design stance, is aware that the strong assumption of optimality cannot be expected to hold in many cases, but he would advocate for it as a first approximation: the initial prediction is that the evolved system conforms to the expectation based on optimality considerations, and then we look for divergences from this prediction. In this way, reverse engineering retains its heuristic value for biology.

However, we might become less sanguine about the value of this strategy as a heuristic the more we attend to the worry that cases of conformity to the predictions based on human design considerations are likely to be rare—the first approximation is likely to be just too wide off the mark. On signalling networks in living cells, Moss (2012) points to research findings of everything 'cross-talking' to everything else. Such networks are nowhere near the ideal of a hierarchical and near decomposable system. Application of a neat, levelled explanatory framework would only be Procrustean. Both Moss and Nicholson (2019: 115) point to a problem with the wiring diagrams commonly used to represent such networks, based on an analogy with electronic networks, because they lead researchers to underestimate the dynamic nature of these signalling pathways, in comparison with a fixed circuit structure.[24] There is a felt need for better analogies, but perhaps they will not be available for the very reason that human engineered systems—at least when they are intelligible enough to usefully serve as analogies[25]—are too fundamentally different from the evolved ones.

[23] It is interesting that these scientists use the term 'machineries' to refer to biological processes, even when their aim is to draw attention to the limitations of reverse engineering.

[24] 'Perhaps the most significant barrier to appreciating the dynamic, heterogeneous aspect of signaling complexes is the lack of a good analogy from our daily experience. This contributes to a second related problem, our inability to depict such interactions diagrammatically. Indeed, the typical "cartoon" of signaling pathways, with their reassuring arrows and limited number of states could be the real villain' (Mayer, Blinov, and Loew 2009: 6, quoted in Moss 2012: 170).

[25] Bongard and Levin (2021) argue, against Nicholson (2019), that twenty-first-century machines, such as deep convolutional neural networks (DCNNs), do not have these rigid, modular qualities that, according to Nicholson, put limitations on the usefulness of machine analogies for living systems. The problem, though, is that self-organising artefacts like DCNNs do not have the intelligibility of simpler, explicitly designed artefacts, and so their utility as an analogy via which living systems can be explained is contestable.

All of these considerations boil down to a concern about oversimplification. By making the assumption that living systems such as the nervous system have distinct levels of organisation, and by using this to justify levelled frameworks in neuroscientific explanation, the density and complexity of brain interactions are most likely being vastly underestimated. Perhaps this does not matter for a range of predictive and technical purposes, but it does undermine more ambitious claims of level-based theories to be unlocking the riddles of information processing in the brain. Potochnik (2021: 75) states the general worry in a compelling way:

> our adherence to the levels concept in the face of the systematic problems plaguing it amounts to a failure to recognize structure we're imposing on the world, to instead mistake this as structure we are reading off the world. Attachment to the concept of levels of organization has, I think, contributed to underestimation of the complexity and variability of our world, including the significance of causal interaction across scales. This has also inhibited our ability to see limitations to our heuristic and to imagine other contrasting heuristics, heuristics that may bear more in common with what our world turns out to actually be like.

The prospect of alternative heuristics is the loaded question. It could well be that the oversimplifications imposed by artefact analogies and level frameworks are indispensable for making such complex biological systems intelligible to human scientists, given our finite cognitive capacities. In which case, there may be no 'better' heuristics, because any attempts to get closer to the actual complexity of the targets result in a loss of tractability and intelligibility. In which case researchers can, without condemnation, settle for the heuristics that they have, but they should uncouple advocacy of their modest explanatory utility from any metaphysical claims about the existence of levels in nature, claims which have been poorly suppressed in previous applications of these frameworks.

4 Concluding Thoughts: The Intransigence of Anthropomorphism

In this chapter I have discussed the most widely known account of levels of explanation in cognitive neuroscience, Marr's three-level framework, and argued that it is not at all easy to separate it from the background ontological commitment to there being separable levels of organisation in evolved systems. I have discussed how the levels framework relies on various analogies with artefacts and design processes, and how such comparisons suggest ways to simplify the brain. Yet, critical concerns can be raised about the framework to the extent that there are only loose analogies and not tight similarities between the brain and the artefacts used to motivate the idea of levels of organization. The ambition is that computational explanations of cognition need not attend to most of the detailed workings of neural tissue, but it could well be that low-level implementational details are crucial to the explanation of high-level cognitive functions. It is likely that the levels framework risks oversimplification of the brain.

In defence of the levels framework it might be said that it is a useful tool, but like any tool it can be misused, and oversimplification is only a risk that comes with misuse. This would be a fine approach if it were possible to understand and deploy the levels framework in only an epistemic manner, as a convenient abstraction that does not commit you to a levelled ontology. However, we should recall here a point made early on in the chapter, that the term *levels of explanation* has both an epistemic and an ontological sense, and also that investigators purporting to take levels in the purely epistemic sense of *levels of abstraction* find themselves creeping into the metaphysical territory of *levels of organization*. If explanation is indeed double faced, and the practice of forming explanations is bound up with taking your targets to be a certain way, then this makes sense of the scientists' lack of restraint, and it is reason to think that oversimplification is the result of normal, not aberrant, use of this conceptual tool.

At the start of this chapter I mentioned Aristotle's anti-reductionist natural philosophy. In his hylomorphic theory of natural substances such as organisms, *form* (which we might equate with structure, organisation, and functionality) takes explanatory precedence over *matter* (the material stuff that a body is made out of). In his presentation of hylomorphism, Aristotle draws analogies with artefacts such as statues and houses to illustrate the difference between form and matter: a completed statue has its characteristic form and matter, but prior to the molding of the bronze there was only matter without the form. For most of the things that people make, one design can be used with different materials, and the same raw materials can be made into different objects, within limits. The same shaped statue of a discus thrower could be executed in bronze or stone or wood, and these materials used to make swords, bridges, and tables. Of course, the design for a bridge would have to be modified, depending on whether it is to be built from wood or stone, given the difference in weight, strength, and flexibility of these materials. But still, the general point holds that with artefacts there is a relative independence of matter and form: the nature of the raw material does not pre-specify what will be made from it, and one design can usually be realised in more than one kind of material.

I think that when we get too used to artefact-organism analogies, we import this intuition into biology—that 'form' is relatively independent of 'matter', and vice versa—and perhaps this is at the root of the attraction of the levels framework as a simplifying strategy in the sciences of brain and mind. For it fosters the expectation that the overall functionality of the system can be understood independently of characterisation of the material constitution. This is a comforting prospect for neuroscientists, like Marr, who recognise the intractability of attempting to derive explanations of how the brain as a whole gives rise to cognitive capacities from models of individual neurons and their interactions. These ideas are also congenial to philosophers who reject reductionist solutions to the mind-body problem. As noted in Section 1, the functionalist 'received view' of Fodor and others took the mind to reside at a computational level autonomous from the level of neural realisation. This offered a route to non-reductive physicalism: creatures with minds are constituted by nothing over and above ordinary matter, but their minds are not identical with nor reducible to states of the nervous system.

The sticking point is that this strong notion of top-level autonomy comes from taking the artefact analogy too far. In Aristotle's own philosophy a disanalogy is noted between artefacts and natural substances. With artefacts there is a contingency—an in-principle multiple realisability, so to speak—of form and matter, but in living bodies form and matter are not held to be just contingently related.[26] This chimes with the viewpoint I advocated in Section 3, that cognitive properties seem to be enmeshed even in the low-level cellular and chemical processes of the brain.[27] While it is clear also that degeneracy and robustness are characteristic of living systems—for example the same developmental pathway can be dependent on a range of different genetic processes, and the brain can regain function following injury by using different neural tissue to perform the previous task—this should not be equated with the multiple realisability posited by functionalism. For the functionalist asserts that cognitive capacities are substrate-independent and could be realised in systems with radically different material bases, so long as certain relationships (e.g., ones defining a computation) are preserved.

Debates in late twentieth-century philosophy of mind took functionalism to be the only alternative to mind-brain reductionism. But by considering Aristotle's anti-reductionism we can see that it is possible to reject the reductionist identification of mind with its material constituent parts, while also rejecting functionalism and its assertion that mind is multiply realisable and autonomous from neuronal material, thought of as merely the implementer of mental programmes. We can also see that reductionism and functionalism—taken as methodological rather than ontological positions—are characterised by two different bets on how to find simplicity, and therefore intelligibility, in the complexity of the brain. The reductionist's hope is that via knowledge of the simpler, small-scale components of the nervous system, studied in isolation from the whole, explanations of global

[26] This is Burnyeat's (1992) interpretation, against views that cast Aristotle as a proto-functionalist. See also Whiting (1992) for a different interpretation which still upholds the relevant difference between natural substances and artefacts.

[27] This gives some indication that form and matter (function and structure) are not contingently related in the brain. A general argument for taking this to be the case in living systems comes from the observation that living things are not assembled from pre-existing components but make their own material components (cells, tissues) in the process of development and self-repair, and these parts are therefore 'tailor made' to serve the functions demanded of them, in likelihood not replaceable by structurally different ones.

cognitive capacities will come about. The functionalist's bet is that the examination of organisation independent of study of the small-scale components will yield explanations of those capacities. But if the alternative position inspired by Aristotle is correct, we see there is a true difficulty in the task of explaining brain and cognition, due to the mutual dependency between small-scale material parts and overall function, meaning that one cannot be understood in isolation from the other, which is what the simplifying procedures of reductionism and functionalism both try to do. This would be enough to elicit shrieks of despair from many researchers in the field. But at least this position makes clear why it is that their work is so difficult! At most, investigators should attempt more modestly to understand the cognitive capacities of the brain in partial isolation from consideration of detailed neurobiology, while acknowledging that such explanations will inevitably be partial because not all of the relevant factors can feasibly be encompassed.

Otherwise, if the artefact analogy is upheld, and considered to highlight a tight similarity between organisation in organic and man-made systems, then the functionalist dreams for simple, explanatory models at the computational level will not die out. In the end, it is important to think more about the prevalence and legitimacy of these analogies. The use of artefact analogies in neuroscience and the rest of biology means that scientists are conceptualizing these systems through the lens of the human—how people design things, and the objects they have made. This is in its own way an anthropomorphised picture of the natural world. Anthropomorphism was supposed to have been expunged from science with the rise of mechanism and the decline of Aristotle's influence. It is therefore ironic that close attention to Aristotle's hylomorphic theory warns us of a disanalogy between organisms and artefacts, and hence a danger of anthropomorphism, one that is not so obvious from other perspectives. The lesson is that computational explanations of mind and brain run the risk of imposing engineering principles onto evolved systems, which leaves this programme of research itself vulnerable to the charge of anthropomorphic projection.

References

Allen, Colin, Marc Bekoff, and George Lauder (eds.). 1998. *Natures Purposes: Analyses of Function and Design in Biology* (MIT Press: Cambridge MA).

Anderson, James A., and Edward Rosenfeld (eds.). 1998. *Talking Nets: An Oral History of Neural Networks* (MIT Press: Cambridge, MA).

Ballard, Dana. 2015. *Brain Computation as Hierarchical Abstraction* (MIT Press: Cambridge, MA).

Barlow, Horace. 1972. 'Single Units and Sensation: A Neuron Doctrine for Perceptual Psychology?', *Perception*, 1: 371–394.

Bechtel, William, and Leonardo Bich. 2021. 'Grounding Cognition: Heterarchical Control Mechanisms in Biology', *Philosophical Transactions of the Royal Society B*, 376: 20190751.

Bechtel, William, and Robert C. Richardson. 2010. *Discovering Complexity* (MIT Press: Cambridge, MA).

Bongard, Joshua, and Michael Levin. 2021. 'Living Things Are Not (20th Century) Machines: Updating Mechanism Metaphors in Light of the Modern Science of Machine Behavior', *Frontiers in Ecology and Evolution*, 9: 650726.

Breitenbach, Angela. 2014. 'Biological Purposiveness and Analogical Reflection'. In Ina Goy and Eric Watkins (eds.), *Kant's Theory of Biology*, 131–158. (Walter De Gruyter: Berlin).

Bullock, Theodore H., Michael V. L. Bennett, Daniel Johnston, Robert Josephson, Eve Marder, and R. Douglas Field. 2005. 'The Neuron Doctrine, Redux', *Science*, 310: 791–793.

Burnyeat, M. F. 1992. 'Is an Aristotelian Philosophy of Mind Still Credible (A Draft)'. In Martha C. Nussbaum and Amélie Oksenberg Rorty (eds.), *Essays on Aristotle's de Anima*, 15–26. (Oxford University Press: Oxford).

Cao, Rosa. 2014. 'Signaling in the Brain: In Search of Functional Units', *Philosophy of Science*, 81: 891–901.

Chirimuuta, M. 2018. 'Explanation in Computational Neuroscience: Causal and Non-Causal', *British Journal for the Philosophy of Science*, 69: 849–880.

Chirimuuta, M. 2021. 'Your Brain is Like a Computer: Function, Analogy, Simplification'. In Fabrizio Calzavarini and Marco Viola (eds.), *Neural Mechanisms: New Challenges in the Philosophy of Neuroscience*, 235–261. (Springer: Berlin).

Chirimuuta, M. 2024. *The Brain Abstracted: Simplification in the History and Philosophy of Neuroscience* (MIT Press: Cambridge, MA).

Churchland, Patricia Smith, and Terrence J. Sejnowski. 1988. 'Perspectives on Cognitive Science', *Science*, 242: 741–745.

Conrad, Michael. 1989. 'The Brain-Machine Disanalogy', *BioSystems*, 22: 197–213.

Craver, C. F. 2007. *Explaining the Brain* (Oxford University Press: Oxford).

Cummins, Robert. 1975. 'Functional Analysis', *Journal of Philosophy*, 72: 741–765.

Cummins, Robert. 2000. ' "How Does It Work?" versus "What Are the Laws?": Two Conceptions of Psychological Explanation'. In Frank C. Keil and Robert A. Wilson (eds.), *Explanation and Cognition*, 117–145. (Cambridge, MA: MIT Press).

Dennett, Daniel C. 1987. *The Intentional Stance* (MIT Press: Cambridge, MA).

Dennett, Daniel C. 1995. 'Cognitive Science as Reverse Engineering: Several Meanings of "Top-Down" and "Bottom-Up" '. In D. Prawitz, B. Skyrms, and D. Westerståhl (eds.), *Logic, Methodology and Philosophy of Science IX (Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science)*, 680–689. Uppsala, Sweden.

Egan, Frances. 2014. 'How to Think about Mental Content', *Philosophical Studies*, 170: 115–135.

Elber-Dorozko, Lotem, and Oron Shagrir. 2019. 'Computation and Levels in the Cognitive and Neural Sciences'. In Mark Sprevak and Matteo Colombo (eds.), *The Routledge Handbook of the Computational Mind*, 205–222. (Routledge: Abingdon).

Fodor, J. A. 1997. 'Special Sciences: Still Autonomous After All These Years', *Philosophical Perspectives*, 11: 149–163.

Godfrey-Smith, P. 2016. 'Mind, Matter, and Metabolism', *Journal of Philosophy*, 113: 481–506.

Grant, Seth G. N. 2018. 'Synapse Molecular Complexity and the Plasticity Behaviour Problem', *Brain and Neuroscience Advances*, 2: 1–7.

Gurney, Kevin N. 2009. 'Reverse Engineering the Vertebrate Brain: Methodological Principles for a Biologically Grounded Programme of Cognitive Modelling', *Cognitive Computation*, 1: 29–41.

Haugeland, John. 1978. 'The Nature and Plausibility of Cognitivism', *Behavioral and Brain Sciences*, 2: 215–226.

Hawkins, Jeff, Subutai Ahmad, and Yuwei Cui. 2017. 'A Theory of How Columns in the Neocortex Enable Learning the Structure of the World', *Frontiers in Neural Circuits*, 11.

Illetterati, Luca. 2014. 'Teleological Judgment: Between Technique and Nature'. In Ina Goy and Eric Watkins (eds.), *Kant's Theory of Biology*, 81–98. (Walter De Gruyter: Berlin).

Kant, Immanuel. 1790/1952. *The Critique of Judgement* (Oxford University Press: Oxford).

Kitcher, Patricia. 1988. 'Marr's Computational Theory of Vision', *Philosophy of Science*, 55: 1–24.

Koch, Christof, and Michael Buice. 2015. 'A Biological Imitation Game', *Cell*, 163: 277–280.

Larkum, Matthew Evan. 2022. 'Are Dendrites Conceptually Useful?', *Neuroscience*, 489: 4–14.

Lillicrap, Timothy P., and K. Kording. 2019. 'What Does It Mean to Understand a Neural Network?', https://arxiv.org/abs/1907.06374.

Love, Bradley C. 2021. 'Levels of Biological Plausibility', *Philosophical Transactions of the Royal Society, B*, 376: 20190632.

Mahfoud, Tara. 2021. 'Visions of Unification and Integration: Building Brains and Communities in the European Human Brain Project', *New Media & Society*, 23: 322–343.

Maley, Corey J. 2021. 'The Physicality of Representation', *Synthese*, 199: 14725–14750.

Mante, Valerio, David Sussillo, Krishna V. Shenoy, and William T. Newsome. 2013. 'Context-Dependent Computation by Recurrent Dynamics in Prefrontal Cortex', *Nature*, 503: 78–84.

Marcus, Gary, and Jeremy Freeman. 2015. 'Preface'. In Gary Marcus and Jeremy Freeman (eds.), *The Future of the Brain*, xvii–xix. (Princeton University Press: Princeton, NJ).

Marom, Shimon, Ron Meir, Erez Braun, Asaf Gal, Einat Kermany, and Danny Eytan. 2009. 'On the Precarious Path of Reverse Neuro-Engineering', *Frontiers in Computational Neuroscience*, 3: 5.

Marr, David. 1982. *Vision* (W. H. Freeman: San Francisco).

Mayer, B., M. Blinov, and L. Loew. 2009. 'Molecular Machines or Pleiomorphic Ensembles: Signaling Complexes Revisited', *Journal of Biology*, 8: 81.

Moss, Lenny. 2012. 'Is the Philosophy of Mechanism Philosophy Enough?', *Studies in History & Philosophy of Biological and Biomedical Sciences*, 43: 164–172.

Nicholson, Daniel J. 2019. 'Is the Cell Really a Machine?', *Journal of Theoretical Biology*, 477: 108–126.

Nunziante, Antonio M. 2020. 'Between Laws and Norms. Genesis of the Concept of Organism in Leibniz and in the Early Modern Western Philosophy'. In Andrea Altobrando and Pierfrancesco Biasetti (eds.), *Natural Born Monads*, 11–32. (Walter de Gruyter: Berlin).

Pasnau, Robert. 2011. *Metaphysical Themes: 1274–1671* (Oxford University Press: Oxford).

Piccinini, Gualtiero, and Carl F. Craver. 2011. 'Integrating Psychology and Neuroscience: Functional Analyses as Mechanism Sketches', *Synthese*, 183: 283–311.

Potochnik, Angela. 2021. 'Our World Isn't Organized into Levels'. In Dan Brooks, James DiFrisco, and William C. Wimsatt (eds.), *Levels of Organization in Biology*, 61–76. (MIT Press: Cambridge, MA).

Schierwagen, Andreas. 2012. 'On Reverse Engineering in the Cognitive and Brain Sciences', *Natural Computation*, 11: 141–150.

Shagrir, Oron, and William Bechtel. 2017. 'Marr's Computational Level and Delineating Phenomena'. In David Michael Kaplan (ed.), *Explanation and Integration in Mind and Brain Science*, 190–214. (Oxford University Press: Oxford).

Simon, Herbert. 1962. 'The Architecture of Complexity', *Proceedings of the American Philosophical Society*, 106: 467–482.

Simon, Herbert. 1969. *The Sciences of the Artificial* (MIT Press: Cambridge, MA).

Simon, Herbert. 1973. 'The Organization of Complex systems'. In H. H. Pattee (ed.), *Hierarchy Theory: The Challenge of Complex Systems*, 245–261. (George Braziller: New York).

Whiting, Jennifer. 1992. 'Living Bodies'. In Martha C. Nussbaum and Amélie Oksenberg Rorty (eds.), *Essays on Aristotle's de Anima*, 75–92. (Oxford University Press: Oxford).

Wu, Jianguo. 2013. 'Hierarchy Theory: An Overview'. In Ricardo Rozzi, S. T. A. Pickett, Clare Palmer, Juan J. Armesto, and J. Baird Callicott (eds.), *Linking Ecology and Ethics for a Changing World*, 281–301. (Springer Science: Dordrecht).

Zeigler, Bernard P. 2002. 'The Brain-Machine Disanalogy Revisited', *BioSystems*, 64: 127–140.

9

Messy but Real Levels in the Social Sciences

Harold Kincaid

This chapter surveys some issues about levels in the sciences as they surface in the social and behavioral sciences. However, it is hard to discuss issues in the social sciences without also discussing the issues about levels in science in general. Thus, I will propose a general take on debates about levels which is what I call naturalist and contextualist: naturalist in that conceptual metaphysics cannot trump science and contextualist in that the use and usefulness of level concepts means different things with different plausibility according to scientific context. I argue that the levels concept is neither incoherent or always useful, that many objections to levels such as doubts about downward causation are anti-naturalist conceptual metaphysical approaches that should be rejected, and that levels cannot be replaced with concepts of mechanisms or scale. All this is tied down and expanded with discussions of some empirical social methodological and research areas. Section 1 sets up the general naturalist and contextualist framework, Section 2 applies these to general debates about levels, and Section 3 discusses in detail some social science research.

1 A Naturalist and Contextualist Framework

Naturalism, of course, gets many different formulations, some sufficiently weak that even the most hard-core conceptual analyst turns out to be a naturalist. The naturalist views I am going to apply to the levels debate claim some or all of the following:[1]

- There is no purely a priori knowledge about the empirical world (the status of mathematics is another set of issues).
- There is thus no purely conceptual knowledge about the empirical world.
- Philosophical conceptual analysis based on linguistic intuitions or what “we” would say does not provide knowledge about the empirical world other than about our psychology, and it is not clear it even does that.
- Metaphysical claims about what science can or could never do cannot be established by conceptual philosophical arguments; they are legitimate only if they are grounded in empirical science.

Harold Kincaid, *Messy but Real Levels in the Social Sciences* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Harold Kincaid 2024. DOI: 10.1093/oso/9780192862945.003.0010

- Philosophy of science is continuous with empirical science and is subject to the same broad empirical standards.

None of these ideas are meant to deny that getting clear on concepts is not a useful enterprise or that exploring conceptual possibilities can promote scientific progress. These uses can and have been important for science and arguably are an important part of scientific progress. However, the test in the end is whether conceptual innovations and analysis contribute to a better understanding of the data. To paraphrase Brecht, science comes first and then metaphysics follows on.

So this naturalism is intended to be a strong version of the doctrine. It says more than that empirical facts matter to metaphysics or that all our beliefs, metaphysical and scientific, are interconnected and tied to empirical facts. This kind of weak naturalism is often invoked by work that is doing what I take to be ultimately a priori conceptual metaphysics. The naturalism invoked here is stricter. Metaphysical claims about science that are in conflict with scientific results are to be rejected; metaphysical claims about science with no obvious empirical base are suspect.

Contextualism as I understand it is a close cousin of the naturalism just described. By contextualism I do not mean the project of showing that the necessary and sufficient conditions for "knowledge" involve context. That is the kind of a priori conceptual analysis I want to avoid. Rather, contextualism is a particular version of various ideas about the holism of belief and testing and of pragmatist ideas about assessing theories in terms of specific purposes. The standard slogans include:

- We are never in the position of evaluating all our knowledge at once (Williams 1995).
- Standards for good empirical inference are themselves empirical and often will hold in limited domains; perfectly general rules for inference are hard to get and even if they are found, they are unlikely to do much on their own without substantive background knowledge (Norton 2021).
- The meaning and use of concepts often varies according to context—knowledge and purposes—and thus often do not have neat definitions in terms of necessary and sufficient conditions (Wilson's (2008) work is a wonderful illustration).
- It is quite possible that in some domains the reality we are interested in is messy or "dappled" (Cartwright's term 1999)—natural kinds with systematic part-whole relations, universal laws, and so on may not be the way things work.
- Philosophical metaphysics is only compelling to the extent it is closely tied to empirical scientific research; while conceptual clarification can certainly be beneficial to science (scientists do it all the time), the further away it is from empirical issues the less trustworthy it is.

These claims are epistemic, semantic, and ontological. Since the notion of context is indeed vague, they can threaten triviality. They need not be when applied to specific problems, however, as we shall see.

2 The Framework Applied to Levels

Applied to debates about levels in scientific research, the above framework has multiple implications. I work through some of them in this section.

Naturalism as advocated above has a number of implications about how debates over levels are conducted. Conceptual, metaphysical arguments about levels are commonplace in the literature. Naturalism is skeptical of them. The right question for the naturalist is what role level claims play in the science. That is a descriptive task and not one addressed by common-sense examples. Being descriptive doesn't preclude making normative judgments or require denying that scientists on occasion can be confused. However, normative claims have to be, broadly construed, scientific claims about what practices are empirically supportable.

Contextualism suggests that we be suspicious about very general claims about levels in science. Whether the claim is that appealing to levels is mistaken ((Potchnik and McGill 2012) or that levels can be eliminated for other notions (Bechtel and Craver 2007; Eronen 2015), levels that are invoked in science are likely to be used in different ways, with different meanings, and for different purposes. In slogan form, analyses are likely to be "local." This claim is a complement to the naturalist claim that the role of levels is not to be determined on general conceptual, metaphysical grounds. A series of more concrete applications of these naturalist and contextualist points to debates about levels follows in the rest of this section and then in the third section on social science.

One useful contextualist point is that levels are invoked in different senses and for different purposes. Levels can be used as descriptive or predictive tools with strong ontological claims, as empirical tools, and/or as explanatory in multiple senses (see List, this volume). As I will detail below, "hierarchical" statistical models may have their main motivation in the fact that aggregate data about groups may improve statistical inference—increase precision in the statistical sense—without any commitment to downward causation or other potential metaphysical issues. Levels are also invoked in science without commitment to either mereology or supervenience. When the cognitive sciences talk about higher and lower levels, they often do not claim that there is some part-whole relation. Neurobiological networks are not "parts" of higher-level cognitive functions; they are at a different level of abstraction in some sense that could usefully be spelled out in different contexts.

Thus, all-purpose accounts of levels—either positive or negative—are likely to be inaccurate in various ways. The part-whole, supervenience view is surely plausible in some cases. For typical eukaryotic cells, they are composed entirely of molecules and cell traits are fixed once the total set of cell molecular facts are fixed. Of course, there may be some interesting complexities where this is not exactly right, but as a generalization and as the basis of the extensive research program that is the basis of molecular biology, it is persuasive. Pointing out that you cannot generalize this to the molecules, cells, organs, etc. hierarchy, as Potchnik and McGill (2012) persuasively do, does not detract from the usefulness of the part-whole, supervenience story in the cell case; once (or when or if) you can describe all the molecular facts about a cell, then other biological characteristics are set. One can agree with Ladyman and Ross (2007) that the picture of the world that sees it made up of little things described by fundamental physics is confused and still think there are other places where the idea of levels makes some sense. It is one reason that multiple papers in this volume talk about levels of facts, not entities.

Trying to replace every use of levels with some other notions is equally suspect. Potchnik and McGill (2012) want to talk of spatial and time scales. Yet cognitive functions and neurobiological networks seem to fit either of these. As we will see, levels in the social sciences do not fit easily into this straight jacket. Replacing levels with mechanisms (Bechtel and Craver 2007; Eronen 2015) seems to be unhelpfully circular when the mechanisms invoked themselves have causal levels—entities, activities, and their *components*.

A related contextualist (and naturalist) point concerns the notion of supervenience. Debates about whether things of kind X depend ("depend" is usually meant ontologically, not casually—but this is a tricky issue) on things of kind Y are very abstract and likely to lead to unproductive conceptual analysis and metaphysics. There are no doubt indefinitely many facts about any one level that might be supposed to fix again probably indefinitely many facts about some other level. Kim (1993) is a classic example, but this general way of proceeding is ingrained in the philosophical discourse. Talk of "properties" fixing other "properties" in the abstract seems like a recipe for vagueness and confusion. The issues are much more tractable if we are talking about the predicates in a specific theory fixing the facts in another theory—e.g., all the facts about individual preferences and endowments fixing all the macroeconomic facts as described by some general account of macroeconomics. The position I advocate wants to look at supervenience claims in terms of well-defined theories or causal accounts and the relations between them.

Thus, to find useful or rejectable supervenience claims we need to get more specific. Hellman and Thompson (1975; 1977) argued early on that it is more useful talking about specific theories with a defined vocabulary fixing the facts of some other theory, but unfortunately much more vague conceptions of supervenience—"properties of kind X" are supervenient on "properties of kind Y"—took over the debates. Instead, focusing on theory-specific determination relations requires knowledge and work that cannot be done by abstract conceptual analysis of "properties" at underdefined "levels." Concrete examples in the social sciences will show up in Section 3.

A central naturalist implication about the levels debate concerns downward causation. There is an entire industry of conceptual arguments around the topic that I am not going to try to cover. However, I can sketch a general perspective, which I will then give more flesh to in talking about the social sciences in Section 3.

The naturalist position I advocate says that what causation is and what can cause what is an empirical question. A priori conceptual arguments are trumped by scientific considerations. That does not mean they have no weight or that scientists cannot be confused, but the burden of proof rests on those who want to overrule scientific claims about causation. There is a large, well-developed literature in the social sciences and in biology, especially ecology and educational research, about the mechanics of multilevel inference (O'Connell and McCoach 2008; Qian et al. 2010; Silvia et al. 2020; O'Connell et al. 2022; Huang 2023). Multilevel inference involves developing models that have base-level units—individuals of some kind—and higher-level facts. The latter can be simple aggregative sums of individual traits, but they can also be characteristics that are in an intuitive sense at a higher level but not describable in terms of properties at the lower level.

Multilevel models can perform two functions: improving descriptive statistical inference and developing multilevel causal claims. The descriptive statistical functions are important because standard regression techniques assume independent observations to calculate standard errors and thus significance values. When observations are instead clustered—as in the case of students all in the same school—then standard errors will be biased downward, increasing type-one mistakes. Adding a factor for such clustering is a common reaction to these problems—basically one controls for the cluster effect. So, including levels has an important epistemic role. However, aside from this epistemic function, multilevel models can be used to make causal inferences from the characteristics of higher-level entities and processes to their causal effect on lower-level entities and processes.

Multilevel causal models can be given a very clear causal semantics using directed acyclic graphs. The graphs then can be instantiated and tested in structural equation models. To show that this is not just a conceptual possibility, I generate a data set with higher- and lower-level entities with specified causal relations. This is a simulation where the regression parameters are set with independent errors and a large set of data based on these elements is produced. This procedure is a standard way of testing whether a given statistical procedure works—data with known characteristics are generated and then the question becomes whether a given procedure produces a close approximation of the known data-generating process. The graph of the data-generating process that I use is in Figure 9.1.

The results from a structural equation model estimating the relations is shown in Table 9.1.[2] The structural equation model is a set of equations representing each

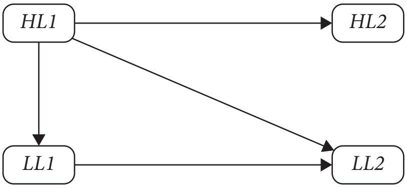

Figure 9.1 Multilevel causal models used to generate data.

Table 9.1 Structural equation results on data generated according to Figure 9.1

Overall model fit statistics:				
Comparative Fit Index (CFI)	0.911			
Tucker-Lewis Index (TLI)	0.733			
Akaike (AIC)	10101.938			
Bayesian (BIC)	10136.292			
RMSEA	0.382			
Regressions:				
	Estimate	Std.Err	z-value	P(>|z|)
hl2 ~				
hl1	1.116	0.031	36.254	0.000
ll1 ~				
hl2	0.610	0.032	18.919	0.000
ll2 ~				
ll1	1.246	0.026	47.879	0.000
hl1	1.125	0.046	24.583	0.000

of the causal relations—in this case four equations capturing each of the causal arrows. HL1 and HL2 are higher-level entities; LL1 and LL2 are lower-level entities. Not surprisingly, there is good statistical model fit for the causal relations in the model.

What the structural equation results show is that a model based on Figure 9.1 fits the data very well. Model fit statistics judge how well the model fits as a whole—maximum likelihood estimation or related methods used to test such models ask how likely we are to see this total set of parameterized relations in this data by chance. The regression coefficients are local tests of the specific relations between variables. All the statistics show very good fit with the simulated data.

Such models are tested frequently in the social and biomedical sciences (Silva et al. 2020). These models claim that casual relations between higher- and lower-level factors are supported or rejected. The question then is what allows philosophers (e.g., Bechtel and Craver 2007) to pronounce that these well-established empirical results are mistaken. My claim is that you should bet on the science, not the philosophical conceptual objections.

Furthermore, a second, fairly simple argument shows that downward causation need not be conceptually confused (Kincaid 1996; see also Yates, this volume). If we are looking at causal claims with a temporal structure (which is the clearest causal case), then we have a description of some whole *W* at time t1 that causes some lower-level element *I* at t2. Even if *W* is simply an aggregate of *I*s, where is the incoherence in saying that the state of *W* at t1 determines the state of some *I* at t2? The cell has a pH of 8 at t1 and that leads at t2 to lower binding levels of proteins *q* and *z*. The inflation rate of an economy at t1 leads individual *I* at t2 to switch from bonds to equities. These are intuitively higher and lower lever elements. It is hard to see anything "spooky" (Bechtel and Craver 2007) here.

There are, of course, conceptual philosophical arguments that higher-level entities and their properties cannot be causes because they are in some sense not real. We will look at some of these in regard to the social sciences in the next section.

As pointed out earlier, science can invoke levels for evidential reasons without commitment to entities with properties at higher levels. "Multilevel modeling" involves a series of general statistical tests that provide ways of dealing with clustered data (see, for example, Heck 2012). Standard multiple regression methods assume that the error terms for each observed individual are independently and identically distributed. If the errors across individuals are correlated, then standard errors are downwardly biased and regression coefficients will be found statistically significant when they are not. Using information about how basic or lower-level entities are clustered—where they are in groups where they share common causes, etc.—allows us to correct such biases. This correction does not require commitment to high-level entities, but ontologically neutral "higher-level" information is needed. So, for example, in time series data the basic units over time will not have independent errors. Each individual forms its own group in that sense over time, and those correlations have to be taken into account to get accurate standard errors and p values. The "group" here at the higher level is not an entity used to casually explain but a higher-level evidential factor, as I noted earlier.

In many cases, multilevel models combine both intra-level and inter-level causation and also make adjustments for clustering, thus combining the causal and epistemic uses of levels. The lack of independent observations because of clustering—the students are all in the same school or classroom—and the school characteristics, e.g., little teacher participation in determining curriculum and the influence on student outcomes, are both at issue. This is probably best done with multilevel structural equation models—causal models with multiple levels and with statistical adjustment for individual groupings where error terms on individual data are not independent (e.g., Heck and Thomas 2020). These points will be fleshed out further in the next section on levels in the social sciences.

3 Levels and the Social Sciences

I argue for four basic points in this section. First, various objections to the idea that social entities can be real causes are mistaken. Seeing this requires a bit of Dennett (1991), Ross (2004), and revealed preference theory. Second, what count as lower levels and higher levels in social research has to be spelled out according to context. The idea of a "supervenience base" is quite fraught and simple part-whole pictures are implausible. Third, there is plenty of solid causal evidence that social entities and their properties can be causes of individual behavior and thus there is evidence for the usefulness of levels in social explanation. I illustrate these three points with some detailed work on the explanation of educational outcomes.

I start with a description of a major tradition in research on educational outcomes. This is just a sketch, but it will be enough to make the points I want to make. I draw the morals about levels afterwards.

A major part of educational research—and the largest literature is about the US educational system—concerns educational "outcomes." These outcomes are usually individual measures based most frequently on test scores of some type: math, reading, etc. To determine what factors explain outcomes, typically some kind of multiple regression techniques have been used. Then deciding what variables go on the right-hand side becomes the major debate, as well as statistical issues about producing reliable results.

Early work looked at student characteristics and did simple OLS multiple regression.[3] Student characteristics included other test scores such as IQ and then traits such as sex, race, social economic status, parents' education, family size, home educational resources, parental occupation, and so on. Note that some of the latter are traits of individual students only in a social relational sense. Being a member of a race, for example, presupposes the whole history and social structure involved in making and sustaining racial distinctions. Most of the early work in this vein did not deal with the nonindependence of error terms resulting from group membership, from being in the same classroom, school, neighborhood, etc., and thus may have produced statistically biased results.

Over time more sophisticated work developed that statistically allowed for clustering by groups, and, especially important for the argument here, allowed for *causal* models with entities seemingly at different levels tested by structural equation modeling and related methods. The variety of nonindividual entities with causal influence is quite interesting. Here is a list of nonindividual variables that have been used in outcomes research:

- Neighborhood
- School district
- Geographical region
- School administrative policies
- Average socio-economic status in school
- School material resources
- School autonomy
- Teacher support
- Level of parental involvement
- Instructional practice
- Classroom climate
- Private school strategies
- Structure of market demand for private school quality

These variables reflect traits of social entities—of classrooms, schools, neighborhoods, districts, and regions. There can be models where there are both causal relations between these social entities—school traits affect classroom traits—and between these variables and individual student outcome measures. Explicit multilevel statistical models capturing these different relationships are possible and increasingly standard (O'Connell and McCoach 2008; Mehta and Petscher 2016; Bardach et al. 2020). These are largely based on observational data, but there are also data from school and other level experimental interventions (Hall and Malenberg 2020).

So, with these examples in hand, let's turn now to issues about levels in the social sciences. I start with skeptical thoughts raised in the literature about social entities as a way to deny levels in the social sciences: if there are no social entities, then there is only one level, viz. that of the individual. There is a long history of ridiculing social entities and factors as real things (Popper 1966). Hegel's "world spirit" has been a prime example, though the notion was probably a bit more subtle than it has been made to be. However, I think there are quite good reasons to think that macrosociological entities can be real causal factors. A main doubt about them is that such entities cannot be real social actors. That argument presupposes that any account of social entities as causal factors has to treat them as "agents" and that treating them so is implausible. Both claims are wrong.

Social entities can have various traits that are not a mere sum of the traits of their members and that are causal factors in their relations with other social entities without treating them as collective *agents*. I detailed a variety of these macrosociological connections some time ago (Kincaid 1996). Governmental traits and traits of firms influence each other's behavior. The inflation rate and the unemployment rate mutually interact. In our education example, there are numerous such causal relations—school traits influence classrooms, district educational organization characteristics influence schools, neighborhood traits influence schools, and so on. And in all these cases, there is the prospect—really inevitability—that these social entities influence individual behavior. None of these causal claims require treating the social entities that interact and influence individuals as if they were *agents* with goals, purposes, etc.

This said, I think there are reasonable ways to treat social entities as collective actors (Kincaid 2019). There are various heroic attempts to argue that collective agents have intentions, purposes, beliefs, etc. in standard folk psychological senses where these are underlying psychological states. List and Pettit (2011) is a good example. Such approaches are extensions of traditional philosophy of action conceptual analysis. That approach, i.e., both the traditional philosophy of action and its attempted extension, seems to me unpromising.

However, an alternative exists that is both grounded in Dennett's notion of "real patterns" (1991) and extended by the microeconomic notion of revealed preference theory developed by Ross (2005). The basic idea is that we can look for consistent patterns in behavior, where "consistent" means the quite tight axioms of revealed preference—transitivity, non satiation, etc.—and then use those patterns as predictors and explainers of further behavior. So, in microeconomics, a prime example is consumer and firm choice. If consumers and firms are consistent in that they fit the tight set of axioms of revealed preference theory, then they can be represented by a utility function for the consumer and by a production function for the firm. Those functions will then predict how they make further choices in the face of, for example, price changes. Agents are being invoked here without attempting to describe the underlying psychological mechanisms.

In the social sciences these kinds of revealed preference or intentional stance models have been widely applied to collective—social—entities. In economics the obvious examples are firms and household, both of which are treated as if they maximize utility functions. In political science, political parties and states are further cases of this approach. In terms of our education example, this approach has been applied to competition between private schools. For example, Bau (2019) models private schools as maximizing entities much as firms are modeled in the industrial organization literature. The models include the choice of school quality produced and purchase price, given characteristics of the market. The model is then estimated with data from Pakistan's rural private schooling market. Schools as social entities interact with each other and have influence on student outcomes and choices.

The point of all the above discussion is that there is good reason to think there is evidence for social levels in a robust causal sense. Social entities are not just epiphenomena floating above the activity of individuals any more than the traits of cells and organelles are epiphenomena above molecules.

The education example also shows that how and whether we talk about levels depends on the factors we want to involve and for what purpose. Clearly, classrooms are composed of students and teachers, schools of classrooms, and school districts of schools. In that sense, there are traditional compositional levels. For the epistemological purposes of avoiding clustering biases, these levels must be considered, and they are in the multilevel statistical models used in educational research as I noted above.

However, while the students and teachers > classrooms > schools > districts composition seems to instantiate the traditional compositional sense of levels, things are not nearly this simple. Instead, there are many factors involved in explaining educational outcomes that are in some sense at different levels but not in any clear compositional sense. So, consider the following:

- Schools are not just collections of classrooms but also involve administration, administrative policies, physical resources, ethnic composition, average social economic status, level of parental involvement, and so on.
- Classrooms are not just teachers and students but teaching practices—curriculum design, instructional methods, etc.—which involve a variety of elements.
- School districts are not just aggregations of schools but also involve neighborhoods, ethnic groups, SES, etc.

These factors do not easily fit into the standard hierarchical picture of levels being "composed of" entities lower down. It is also unclear what would count as a supervenience base. I think it is an empirical matter that varies from context to context and question to question whether and what kind of "levels" and relations between them we need to explain educational outcomes.

Also, the implications for "downward causation" in the social sciences are fairly obvious. There are multiple senses of levels, and the causation between them has to be evaluated by specifying what sense of levels we have in mind. Assuming that classroom traits are entirely determined by traits of students and of teachers, this notion of supervenience still allows for the dynamic sense of downward causation described above: the total state of these variables at one time causally influences student outcomes at another. Moreover, the complexities described above about levels make for even more uncontroversial claims about causation between levels. Suppose the claim is that the physical and material resources of a school—an aggregate characteristic—causally influence student outcomes, an outcome at the component level. What is the objection to this claim about downward causation between levels? It is hard to see what it might be, barring a longer metaphysical story.

4. Conclusion

My hope is to have made at least a prima facie case that talking about levels in the social sciences is not incoherent but certainly in need of clarification. Philosophers can contribute, but that clarification however has to be closely tied to social science research. Levels in the social sciences come to different things in different investigative circumstances. Sometimes they support the traditional compositional hierarchy, often they do not. Yet the concept of levels remains useful in understanding social science practice, albeit contextualized for the explanatory, theoretical, and evidential goals at issue.

Notes

[1] See, for starters, Quine (1969, 1980), Williams (1995), Wilson (2008), and Maddy (2007).

[2] The data were simulated with TETRAD (Glymour et al. 1988) and the SEM test was done with the lavaan package in R (Rosseel 2012).

[3] Good references here that I draw on are the essays in O'Connell and McCoach (2008) and especially Ma et al. (2008) and Bardach (2020).

References

Bardach, L., Yanagida, T., and Lüftenegger, M. 2020. Studying classroom climate effects in the context of multi-level structural equation modelling: an application focused theoretical discussion and empirical demonstration. *International Journal of Research & Method in Education*, 43(4): 348–363.

Bau, N. 2019. *Estimating an Equilibrium Model of Horizontal Competition in Economics. Discussion paper DP13924*. Center for Economic Policy Research.

Bechtel, B., and Craver, C. 2007. Top-down causation without top-down causes. *Biology and Philosophy* 22: 547–563.

Cartwright, N. 1999. *The Dappled World*. Cambridge: Cambridge University Press.

Dennett, D. 1991. Real patterns. *Journal of Philosophy* 88: 27–51.

Eronen, M. 2015. Levels of organization: a deflationary account. *Biology and Philosophy* 30: 39–58.

Glymour, C., Scheines, R., Spirtes, P., and Kelly, K. 1988. TETRAD: discovering causal structure. *Multivariate Behavioral Research* 23(2): 279–280.

Hall, J., and Malenberg, L. 2020. The contribution of multilevel structural equation modelling to contemporary trends in educational research. *International Journal of Research & Method in Education* 43(4): 339–347.

Heck, R. 2012. Multilevel modeling with SEM. In Marcoulides, G., and Schumaker, R. (eds.), *New Developments and Techniques in Structural Equation Modeling*, 895–908. Psychology Press: New York.

Heck, R., and Thomas, S. 2020. *An Introduction to Multilevel Modeling Techniques*. New York: Routledge.

Hellman, G., and Thompson, F. 1975. Physicalism: ontology, determination and reduction. *Journal of Philosophy* 72: 551–564.

Hellman, G., and Thompson, F. 1977. Physicalist materialism. *Noûs* 11: 309–345.

Huang, F. 2023. *Practical Multilevel Modeling Using R*. Sage: Thousand Oaks.

Kim, J. 1993. *Supervenience and Mind*. New York: Cambridge University Press.

Kincaid, H. 1996. *Philosophical Foundations of the Social Sciences*. Cambridge: Cambridge University Press.

Kincaid, H. 2019. Causalism and acausalism. In Schuman, G. (ed.), *Explanation in Action Theory and Historiography Causal and Teleological Approaches*, 179–194. New York: Routledge.

Ladyman, J., and Ross, D. 2007. *Every Thing Must Go*. Oxford: Oxford University Press.

List, C., and Pettit, P. 2011. *Group Agency*. New York: Oxford University Press.

Ma, X., Ma, L., and Bradly, K. 2008. Using multilevel modeling to investigate school effects. In O'Connell, A, and McCoach, B. (eds.), *Multilevel Modeling of Educational Data*, 59–110. Information Age Publishing.

Maddy, P. 2007. *Second Philosophy: A Naturalistic Method*. Oxford: Clarendon.

Mehta, P., and Petscher, Y. 2016. Using n-level structural equation models for causal modeling in fully nested, partially nested, and cross-classified randomized controlled trials. In Harring, J., Stapleton, L., and Baretvas, S. (eds), *Advances in Multilevel Modeling for Education Research,* 193–228. Information Age Publishing: Charlotte, NC.

Norton, J. 2021. *The Material Theory of Induction*. Calgary: University of Calgary Press.

O'Connell, A., McCoach, B., and Bell, B. 2022. *Multilevel Level Modelling Methods*. New York: Information Age.

O'Connell, A., and McCoach, D. 2008. *Multilevel Modelling of Education Data*. Charlotte, NC: Information Age Publishing.

Popper, K. 1966. *The Open Society and Its Enemies*. London: Routledge & Kegan Paul.

Potchnik, A., and McGill, B. 2012. The limitations of hierarchical organization. *Philosophy of Science*, 79: 120–140.

Qian, S., Cuffney, T., Alameddine, I., McMahon, G., and Reckhow, K. 2010. On the application of multilevel modeling in environmental and ecological studies. *Ecology* 91(2): 355–361.

Quine, W. V. O. 1969. *Ontological Relativity and Other Essays*. New York: Columbia University Press.

Quine, W. V. O. 1980. *From a Logical Point of View*. Cambridge: Harvard University Press.

Ross, D. 2004. Rainforest realism: a Dennettian theory of existence. In Ross, D., Brooks, A., and Thompson, D. (eds.), *Dennett's Philosophy: A Comprehensive Assessment*. MIT Press, 147–168.

Ross, D. 2005. *Economic Theory and Cognitive Science: Microexplanation*. MIT Press.

Rosseel Y. 2012. lavaan: an R package for structural equation modeling. *Journal of Statistical Software*, 48(2): 1–36.

Silva, B., Bosamcianu, C., and Levente, L. 2020. *Multilevel Structural Equation Modeling (Quantitative Applications in the Social Sciences)*. Thousand Oaks: SAGE

Williams, M. 1995. *Unnatural Doubts*. Princeton: Princeton University Press.

Wilson, M. 2008. *Wandering Significance*. Oxford: Oxford University Press.

PART IV
LEVELS OF EXPLANATION IN PHYSICS

10

Levels Worth Having

A View from Physics

Eleanor Knox

1 Introduction

Our world appears to be well described by a richly layered set of theories and models. The full scientific picture requires virology and organic chemistry as well as quantum field theory and general relativity. And virology is not a coarse-grained, zoomed-out, or aggregated version of some fundamental physics theory. But it was not always obvious that this was the case. Imagine offering an account of levels of description or explanation from the perspective of Philosophus Paleas, a mid-nineteenth-century philosopher of physics. Paleas observed that widespread application of Newtonian mechanics had been wildly successful at scales from the atomic to the astronomical. Granted, there was often much subtlety in the application of Newtonian dynamics at a level—the kinetic theory of gases required hard theoretical work. But Paleas might reasonably have expected Newtonian dynamics to have interesting applications at more or less every scale; offering a new Newtonian level of description seemed to be simply a matter of zooming out, and choosing an appropriate grain such that, for example, centres of mass could be treated as New- tonian point particles. Masses and forces could be aggregated to yield descriptions in terms of the same kinds of quantity at each scale. The physics imagined by our fictional philosopher has levels in a sense—describing a gaseous planet in astronomical terms is very different from treating it as a ball of ideal gas—but these levels are, in a sense that I will explain here, cheap and plentitudinous. While some choices of scale or grain may be more interesting than others, any choice might offer some kind of physical description.

Our world is not like that. Indeed, we now understand that it could not possibly have been as Paleas imagined.[1] A world describable by Newtonian mechanics

[1] Needless to say, this is all fiction. Paleas is Latin for straw, and it seems doubtful that anyone in the nineteenth century would have described themselves as a philosopher of physics (although Ernst Mach certainly could have). But I think that certain metaphysically fundamentalist, reductionist views have more in common with our fictional philosopher than they may realise.

Eleanor Knox, *Levels Worth Having* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Eleanor Knox 2024. DOI: 10.1093/oso/9780192862945.003.0011

at all scales has neither stable atoms for the kinetic theory to describe nor the coarse-grained astronomical structure necessary to model, say, our solar system. Within physics (let alone elsewhere), our stable of theories holds a vast array of dynamical modelling techniques from the quantum field theories of the standard model to the equations of fluid dynamics, to the cosmological implications of general relativity. Although we often have the inter-theoretic relations in play under reasonably good control, the equations of these theories bear little resemblance to one another and apply at specific scales and in specific domains; one cannot smoothly change one's scale and continue to use, say, the characteristic mathematics of fluid dynamics.

This essay will explore the ways in which theories must interact to make such a thing possible—why is it that we have distinct levels of description and explanation that involve very different dynamics and variables? In physics, one must answer this question with an eye to the fact that we often have a good mathematical handle on how two levels of description are related. Philosophers of physics thus often espouse a view of 'emergence' that is compatible with reduction. Work by Jeremy Butterfield (2011a, 2011b) has convinced many philosophers of physics that much interesting physics behaviour is emergent in the sense of being novel and robust, while also being reducible.

I'll argue here that theoretical descriptions that are both novel and robust give us (partial, local) levels worth having. Novelty is a key ingredient which sometimes goes under-analysed in favour of robustness (which I'll argue here is closely related to autonomy). But novelty is crucial if we are to have sparser, higher-value levels, rather than the cheaper plenitudinous sort our Newtonian philosopher imagined. Alex Franklin and I have argued (Knox 2016, 2017; Franklin and Knox 2018) that certain kinds of variable changes allow for novel explanations. I'll argue here that this kind of novelty gives us levels worth having.

This debate plays out somewhat differently in physics than it may elsewhere (although I think that physics examples should be instructive in the non-physics special sciences). In Section 2, I'll discuss the kinds of levels I take myself to be talking about; these are highly local, and often system-specific. In Section 3, I'll look at what it takes for a level of description to be autonomous or robust. I'll argue that, while these notions are crucial, more is needed if we want an interesting and sparse account of levels. In Section 4, I'll offer an account of novelty that I think can return such an account and suggest that it applies in at least some cases outside of physics.

2 What kind of levels?

The first two chapters of this book present some far-reaching thoughts about levels. Christian List describes a very general programme for understanding levels structure.[2] List describes levels as a set of objects connected by mappings with particular formal properties. The programme is flexible enough to incorporate a wide variety of views on levels but is best suited to widespread or global levels connected by one of the classic inter-level relations such as derivability or supervenience. Angela Potochnik, in contrast, recommends that we give up on (explanatory) levels altogether. Our explanatory strategies are too heterogeneous and their relations too complex to admit of a neat levels hierarchy.

My approach to levels sits somewhere between these views. I agree with Potochnik that much of the levels literature fails to recognise the opportunism and heterogeneity of our modelling and explanatory strategies. I also agree that part-whole relations cannot underpin a neat levels hierarchy: simple examples in which part-whole relations underlie reduction do not generalise. I nonetheless think that there is a crucial role for levels talk in our attempts to understand the world. That is, in part, because, unlike Potochnik, I am not an anti-reductionist; I think theoretical reduction is often possible. We often understand the relationship between different physical descriptions of a system exceedingly well, and there is a great deal of scientific understanding to be had from the study of inter-theoretic relations. However, I do not think that physics examples fit very well with List's formal programme, even if they can be shoehorned into it. Grand theoretical levels—'the level of particle physics', 'the level of macro-dynamics'—do not get much purchase in physics. It is better to think of levels as applying to our various modelling practices for some given target system.

Consider the example of the sun.[3] A standard textbook on the topic, Dermot Mullan's *Physics of the Sun: A First Course* (Mullan 2022), begins with some simple Newtonian gravitational theory in order to calculate gravitational acceleration and escape velocities at the surface of the sun. It goes on to discuss classical optics (opacity, for example), before moving on to considering atomic spectra. By chapter 4, statistical mechanics and the ionisation properties of gases have taken centre stage. Hydrodynamics is crucial in the convection zone of the sun and occupies several chapters. As you'd expect, modelling nuclear fusion in the interior of the sun takes up several more. At this stage we've reached chapter 11 of 18, and I would imagine the reader gets the picture: describing the physics of the sun requires resources from most corners of physics, each employed to provide a different model, either of the sun itself, or of some part of it.

At first glance, this example looks as if it supports Potochnik's levels scepticism. The study of a single system requires multiple explanations that interact in various cross-cutting relationships. Nonetheless, I think that levels talk is helpful here.[4]

[2] For more on List's view of levels, see List 2019.

[3] Thanks to David Wallace for making me think about this example.

[4] See Alex Franklin's chapter in this volume for a full argument as to why these multiscale explanations need not threaten reduction or levels talk.

This is in part because I am *not* a theoretical anti-reductionist. We have a good grasp on the relations between many of the theories used in the physics of the sun. For example, the relation between discrete atomic and fluid descriptions of matter, while complex and interesting, is well understood. Statistical mechanics just *is* the study of certain aggregated and statistical properties of atoms and molecules and originated in the desire to understand how thermodynamic properties related to smaller scales. Granted, it doesn't make much sense to ask about the sun 'at the level of statistical mechanics', but that is because we have been too vague in our labelling of levels. It makes perfect sense to ask about the sun's corona as modelled by the Saha equation for the ionisation of gases, which is a statistical mechanical model. By contrast, understanding the fusion processes that power the sun requires modelling at the level of nuclear physics, and we understand the relationship between nuclear physics and atomic physics, and between atomic and molecular physics and statistical mechanics, surprisingly well given the complexity of the theories involved. Much (if not all) of our solar description can be connected by local inter-theoretic relations that fit the levels framework.[5]

Why talk about local modelling techniques rather than global theories? In part because I have something like a semantic view of theories in mind,[6] and in part because our theories can apply at vastly different scales. Consider the example we started with, in which Newtonian point particle mechanics could be applied both to the ideal theory of gases and to the dynamics of our solar system. What might it mean to describe the world 'at the level of Newtonian point particle mechanics'? Are we talking about the level at which we treat Jupiter as a Newtonian point particle, or the level at which we treat the molecules of its gases as point particles? If one thinks this kind of level is under-specified at a global scale, then consider how much less clear it might be to talk about 'the level of physics', 'the level of chemistry', or 'the level of biology'.

The picture that emerges is this one: individual systems, small and large, can be modelled in multiple ways. These modelling methods often correspond to theoretical levels, and we often (although not always!) have a very good theoretical understanding of the relationship between the models. We might well want to call the relationship between these levels reduction. In many cases we can find a way of cashing them out in terms of Nagelian reduction if we are sufficiently liberal about the mathematical complexity of our bridge laws. If one feels (as I do) that Nagelian reduction isn't quite the right framework for these kinds of inter-model relationships, one might note that in many cases we can answer all the scientifically

[5] For interesting philosophical accounts of some of the relations involved see, for example, Wilson 2006, Wallace 2021a, or Knox 2016 (for my own views on one relationship involved).

[6] Specifically, something like Wallace's maths first semantic view of theories (Wallace 2021b).

interesting questions about the relationship between models. When there are no further inter-theory mysteries to be investigated, reduction seems to be the relation at hand.[7]

However, it is not just the anti-reductionist à la Potochnik who might be sceptical about levels. Reductionism can also lead to levels-scepticism; if our higher-level physics is reducible, why do we need the higher level at all? This is where the distinction between cheap, plenitudinous levels and sparser, more meaningful, ones comes in. Suppose, in keeping with the fictional philosopher of the introduction, that all physics was Newtonian point particle mechanics. There would still be pragmatic advantages to particular choices of scale, grain, or approximation given particular problems. Point particle mechanics can be applied to point particles, or to planets, or to the balls on a billiard table. If all were Newtonian, we might expect this to apply at more or less any scale and for there to be interesting problems for each of these choices. We might call these levels of description or explanation if we wished. Such cheap and easily come-by levels would not, however, carry much weight when it came to reading off the ontology of the world, or establishing the need for different sciences. If we wish to call these levels, they risk triviality. As a result, the distinction between the believer in plenitudinous levels and the reductionist levels-sceptic might seem a mere matter of perspective; neither put great weight on the importance of non-fundamental levels of description, and both agree that we are free to coarse-grain the fundamental description in various ways.

Those interested in defending ontologically significant levels talk will want something much stronger than the above. I will argue in Section 3 that they can have something stronger, even in the presence of successful reduction—even where we can *explain* why one level of description and explanation holds in terms of the others. The key here, at least within physics, will be that certain changes of variable make clear dependencies that are not revealed by the lower-level choices of variable. This will allow for *explanations* that are novel, even while the relationship to lower-level physics is well understood. Crucially, not all scale changes or changes of grain will do this, and it is only where we have a new set of dependencies revealed by variable choice that we should talk of a novel level of description and explanation.

[7] Many contemporary views on reduction involve some liberalisation of strict Nagelian reduction. These range from slight softenings of Nagel's programme to accounts of reduction that eschew logical derivability as a standard. For example, Dizaji-Bahmani, Frigg, and Hartmann (2010) offer a defence of a liberalised Nagelian framework. Robertson (2022) and Franklin (2019) propose a less Nagelian framework. More needs to be said about what reduction means in a more semantic framework. Wallace (2021b) cashes reductive relationships out in terms of instantiation—models of, say, discrete particle dynamics can instantiate models of fluid dynamics. However, he stops well short of giving a full account of the instantiation relation and its relation to traditional discussions of reduction.

3 Autonomy/robustness

In the broadest possible strokes, one might want to divide the puzzle that stems from the reducibility of physics into two pieces:

(1) Why doesn't fundamental physics matter more?
(2) Why is there anything other than fundamental physics?[8]

As previously mentioned, Jeremy Butterfield (2011a and 2011b) proposes that we identify emergence in physics with novel and robust behaviour, although he then goes on to say much about robustness and very little about novelty. In my view, robustness is crucial to the first question, and novelty to the second, and both are needed to get the kind of weak emergence that allows for interesting levels.

Let us start in this section with the first question. We don't have a theory of fundamental physics. Not only do we not have a unified theory of gravity and the other forces (a theory of quantum gravity), but our most fundamental theory in the non-gravitational domain, quantum field theory, is an effective field theory that breaks down at smaller scales. It is fair to say that, if there is a fundamental theory, we know relatively little (although not nothing) about what kind of theory it might be. And yet science plows on despite slow progress in theoretical physics. Physicists get away without knowing much about more fundamental physics, and biologists get away without knowing much about physics. The details of the micro-physics simply don't matter for a great many purposes. This is the phenomenon of the autonomy of the special sciences, and I take it to be closely related to the phenomenon of robustness.

A feature of a system is robust if it remains unchanged under perturbations of other variables. Generically, approximations and abstractions will yield robust quantities: mass as rounded to the nearest kilogram is robust under variations of mass as measured in grams. But there are more interesting examples of robustness as well. Within biology, there has been much discussion of biological robustness: for example, Hiroaki Kitano defines robustness as 'a property that allows a system to maintain its functions despite internal and external perturbations' (Kitano 2004). Many discussions of asymptotic behaviour in physics can be seen as a demonstration of robustness—for example, Bob Batterman's discussion of the rainbow (2002a and 2002b) falls into this category. Likewise, when one uses the thermodynamic limit to explain phenomena like phase transitions, taking the particle number to infinity, one can see this as a demonstration that the phenomenon is independent of the exact particle number, as long as it is large enough.

[8] It's not a coincidence that these are close to the titles of two pieces of philosophy that I have found very helpful: Alex Franklin's PhD thesis (2019) and a recent article by Katie Robertson (forthcoming).

The examples above demonstrate a specific kind of *autonomy* of the higher level. Katie Robertson (forthcoming) has recently offered a generalised account of autonomy based on Woodward's account (Woodward 2016, 2021) of explanatory autonomy that may be helpful here. According to Robertson, autonomy occurs when:

> . . . a microfact b is unconditionally relevant to the macrofact A, but conditional on macrofact B, b is irrelevant to A. One way of putting this: the macrodependence between A and B *screens off* the microdetails. (Robertson, forthcoming, 9)

Relevance is here defined in terms of probabilities:

- **Unconditional relevance**: Microfact b is unconditionally relevant to macrofact A when: $P(A|b) > P(A)$.
- **Conditional irrelevance**: Microfact b is irrelevant to A conditional on macrofact B when $P(A|B\&b) = P(A|B)$.

The idea here is clear: sometimes it is the case that although some microdetail might be relevant to a given macroscopic fact, there exists some other macroscopic fact that tells us everything we need to know. Once we know macrofact B, we no longer need to appeal to microfact b. Robertson argues that this kind of generalised autonomy can capture various aspects of autonomy in the literature, from the kind of dynamical autonomy that our robustness examples above display, to causal autonomy and nomic autonomy.

This account is helpful. It gives a clear sense of what we mean by the term 'autonomy' that coheres well with much of the literature. It provides an answer to our first question above: fundamental micro-physical details might matter for some phenomenon, but they don't matter once we conditionalise on the macroscopic details that we already know. But Robertson's definition cannot help us to find an account of sparse, high-value levels.[9] Autonomy in this sense is present whenever we use any approximation whatsoever. Consider a car safety engineer explaining why a given car is safe in a crash into another car at 30 miles per hour. She'll appeal to a number of facts relevant to the acceleration experienced by passengers in the car, among them the masses of the two cars involved in the collision. She'll also consider what acceleration average adults can safely handle. The mass of each car and each passenger in grams will be unconditionally relevant to the survival of the car's passengers, but will be irrelevant conditional on their mass rounded to the nearest kilogram—or to the nearest 10 kilograms for that matter; the fact that one of our passengers may have eaten a large turkey dinner before setting off does not

[9] I should be clear here that Robertson doesn't intend it to!

affect the explanation. And yet only on the most minimal account of levels (List's levels of awareness, perhaps) could a change in precision of measurements from grams to kilograms in this kind of setting betoken a new level.

Likewise, the world imagined by our fictional philosopher of physics is one with plenty of autonomy, but very little novelty. The world they imagine is one in which not only might one describe the crash above in units with arbitrary precision, but also one in which one could consider the relevant bodies—the cars and the passengers—in terms of the dynamics of their constituent parts, and simply aggregate the relevant masses and forces. To use a metaphysical picture much derided by contemporary philosophers of physics,[10] our nineteenth-century physicist might imagine the car and its passengers as if built from Lego. The relevant variables that describe the car—mass and velocity, for example—are the same variables that describe the bricks. And the mass of the car is simply the sum of the mass of its constituent Lego bricks. Such a picture leaves much room for autonomy. The masses of the constituents of the car are unconditionally relevant to the survival of the passengers, but conditionally irrelevant given the total mass. If our Lego bricks come from a factory with a slightly variable plastic mass per brick, that won't much change the account: the explanation here is robust under a wide range of perturbations of the smaller mass variables. But description at the Lego scale has many of the same features as description at the car scale. Moreover, one can imagine a sequence of scales between the Lego scale and the car scale, each characterised in much the same way; our choice here of Lego offers natural discrete Lego units, but we could offer descriptions in terms of multiples of these. Such descriptions correspond only to levels in the cheap or plenitudinous sense: they are autonomous but not novel. High-value levels will also require novelty.

Robustness and autonomy are important. Robustness demonstrations, in particular, can be crucial to understanding the dynamical stability of a phenomenon. But they tell us very little about how and why systems admit of descriptions and explanations at different levels. The trouble here is that floating free of precise microphysical details is not enough to make particular scales and levels novel in the way that, say, fluid dynamics seems to be novel as compared to a particle dynamics description, let alone the way that biology seems to be novel relative to particle physics. Such theories are not only autonomous relative to their lower-level counterparts, but also involve radically different properties, dynamics, and variables. In the fluid dynamics case, we move from a discrete dynamics described by quantum mechanics to a continuous description governed by the Navier-Stokes equation. Butterfield was right to say that (weak) emergence requires novelty and robustness, but wrong to push the important notion of novelty into the background.

[10] See, for example, Ladyman and Ross 2007.

4 Novelty from variable changes

So let us turn to the second question. Why is our nineteenth-century philosopher wrong? Why isn't science just Newtonian physics, or, slightly better, some kind of quantum theory, all the way down and all the way up? *Why is there anything other than fundamental physics?* One *could* answer this question via an anti-reductionist picture of levels: there is something other than physics because there really are systems that aren't described by physics. But this sits ill with the kind of system-specific reductionism available to us in many domains. Instead, we should answer the question by pointing out that some levels of description are not merely robust under changes to the micro-physical detail, but *novel* when compared to the micro-physics.

What should we make of novelty? There are fewer accounts in the philosophy of physics literature than one might like. Butterfield defines it in subjective terms: novel features are surprising relative to the lower-level description. A more substantial option, taken in Wilson (forthcoming), is to claim that descriptions are novel when they appeal to novel, higher-level natural kinds. But this account feels like mystery mongering—why are these higher-level properties natural kinds and how do we distinguish real natural kinds from gerrymandered ones? What is it about our world that allows there to be such natural kinds even when reduction is possible? Our account of higher-level natural kinds should itself depend on the details of our theories. In my view, the proponent of higher-level natural kinds needs to appeal to another account of novelty, both to explain our epistemic access to natural kinds, and to explain what they are in the first place.

Katie Robertson and Alex Franklin (2021) aim to explain which kinds we should admit into a richly layered and expansive 'rainforest realism'.[11] They claim that what is importantly novel about interesting levels is that they involve entities that feature in macro-dependencies with a distinct functional form from the corresponding micro-dependencies. Intuitively, there is something right about this; it captures the sense in which our purely Newtonian world fails to have novel levels of description. Moreover 'dependencies' seem to be the right place to focus our attention; new levels of description and explanation often reveal relevance relations (potentially causal ones) that were not accessible at the lower level. And yet 'distinct functional form' might be a little hard to pin down. I think we get a clearer view of novelty (and a better idea of what we might mean by 'distinct functional form') if we think in detail about the relationships between physical variables.

Certain kinds of changes of variable allow us to see relevance relations and dependencies that are opaque until the correct variable is chosen. Consider, for

[11] This is a reference to Ross 2000.

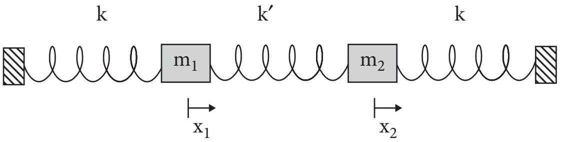

Figure 10.1 Coupled masses on springs.

example, the following simple coupled oscillator.[12] Two particles of equal mass *m* oscillate on springs with constants *k* and *k*', as shown in Figure 10.1:

Their motion is characterised by the following equations:

$$m\ddot{x}_1 = -kx_1 - k'(x_1 - x_2) \tag{1}$$

$$m\ddot{x}_2 = -kx_2 - k'(x_2 - x_1) \tag{2}$$

Coupled second order equations like this are hard to solve for the x_1 and x_2 variables. However, if one makes a simple transformation of variables:

$$\eta_1 = x_1 + x_2$$

$$\eta_2 = x_1 - x_2, \tag{3}$$

one can convert equations (1) and (2) into linear uncoupled differential equations for two simple harmonic oscillators:

$$m\ddot{\eta}_1 = -k\eta_1 \tag{4}$$

$$m\ddot{\eta}_2 = -(k + 2k')\eta_2. \tag{5}$$

Solutions to this kind of equation are taught to every first-year physicist:

$$n_1 = 2A_s \cos\left(\sqrt{\frac{k}{m}t + \phi_s}\right) \tag{6}$$

$$n_2 = 2A_f \cos\left(\sqrt{\frac{k + 2k'}{m}t + \phi_f}\right) \tag{7}$$

[12] This particular example was first brought to my attention by Alex Franklin and appears in Franklin and Knox 2018.

These two equations and their solutions characterise normal modes of the system. One of these corresponds to the two masses oscillating together (so their distance remains constant), and another to the two masses moving in opposite directions. We'll call the variables η_1 and η_2 that define these modes *normal mode variables*.

Suppose we're interested in compression in the central spring, or in some phenomenon that depends on it. Moving to the normal mode variables allows us to understand this directly—it's only the distance between the two masses, represented by η_2, that is relevant; spring compression depends only on η_2. Choosing the right variables brings this dependence into focus.

In the case above, the dependence seems straightforwardly causal—distance between the masses is the cause of the spring compression. James Woodward (2016) explores the effect of variable changes on causal inferences. His focus is the way in which variable changes can make independent variables dependent, and vice versa. This means that choice of variables is crucial for causal analysis. Woodward's goal in this paper is to analyse the messy and crucial process of causal modelling. He aims to find a heuristic for variable choice for researchers and defends a number of criteria that might be employed in statistical analysis. But suppose we already have a successful model to hand, and we understand the relationship between the variables used and those at another level. Then the moral of Woodward's paper is that the higher-level description can reveal causal dependencies that are not revealed by the lower-level description. In our normal modes case, 'mixing' the variables reveals a new set of causal relationships.

What does all of this have to do with levels? Weakly emergent levels as defined here are those whose descriptive variables are novel and robust with respect to some lower level. We now have our finger on an interesting, accessible, and non-subjective kind of novelty. In earlier work (Knox 2016 and Franklin and Knox 2018) I argued that the novelty is explanatory. When we choose the right variables, we reveal a dependence that allows us to give better explanations, because getting the dependence relations right allows us to abstract to exactly the right level of detail. In the case above, choosing the normal mode variables allows us to explain the spring compression in terms of just what is relevant—variable η_2 and not η_1. This is crucial to why we have and need levels of explanation.

However, there is a risk that talk of novel explanatory value sounds too epistemic to ground levels in any interesting sense. It's important here that the dependence relations revealed by variable changes are real dependencies[13]—it's a fact about the world, and not merely our description of it, that only distance between the springs is relevant to spring compression. In past work, I've eschewed talk of causation due

[13] This emphasis on dependence, rather than just explanation, is present in Alex Franklin's work, and I've been persuaded by him that it's helpful.

to Russellian qualms about identifying clearly causal relationships in more fundamental physics theories. But at the levels of description at issue here, the novel dependencies revealed by variable change are causal. The explanations are causal explanations, and novel explanatory power is connected to getting the causal description right.

I am not suggesting that, in the example above, the normal modes description gives us a new level of description in any non-trivial sense. The normal modes variables depend very sensitively on our x_1 and x_2 variables. They are not robust or autonomous with respect to the underlying description. The following thought is tempting: just as the Newtonian world picture of our fictional philosopher offered autonomy without novelty, this relationship gives us novelty without autonomy. But this underplays the complex relationship between autonomy and novelty. I suspect most readers will be reluctant to think that the dependence relations revealed by our move to normal modes are truly novel—after all, distance between the masses feels like a perfectly obvious quantity to examine in our original description. Novelty does not, therefore, float entirely free of robustness and autonomy. While our normal modes example demonstrates the power of variable choice to pinpoint dependencies, we need a more complex relation between variables in order to truly think of these dependencies as 'novel'.

Happily, examples of something like normal modes in more complex systems in physics are easy to come by. Franklin and Knox (2018) examine the example of phonons, which are the normal modes of macroscopic crystals. The move from a description with numbers of atomic displacements on the order of 1,026 to the vibrational modes of the crystal involves many more approximations and idealisations which ultimately result in a powerful and robust description. It's also one that is very closely tied to how particle descriptions relate to the quantum field. But these examples would take us further into the physics weeds than is appropriate to this volume. Instead, I'd like to briefly look at a neuro-physiological example.

The stomatogastric ganglion of a lobster is a simple neuronal network. It consists of 30 neurons located on the wall of the digestive tract of the lobster. It contains two central pattern generators (CPGs): that is, generators of characteristic signals that control basic digestive behaviours. The pyloric CPG controls peristaltic motions that pass food down the gut. Of interest to us here, however, is the gastric CPG, which controls the motion of three internal teeth. The gastric CPG typically generates two different patterns, which move the teeth in different ways: Type 1 patterns cause the three teeth to squeeze together simultaneously and Type 2 patterns cause the two lateral teeth to move in opposition to the medial tooth in a cut and grind motion.

Figure 10.2 shows recordings of muscle contractions in live lobsters. The gastric mill CPG is comprised of 10 motor neurons that stimulate muscles, and just one connecting neuron, so muscle contraction is an accurate proxy for groups of neurons firing. The top two signals show the muscles that protract and retract the

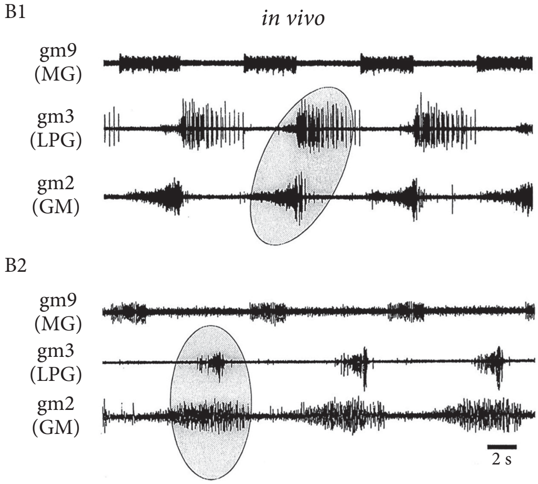

Figure 10.2 Gastric Mill output patterns as measured in live lobsters. Reproduced from (Coombes et al. 2002, 583). This image is not covered by the terms of the Creative Commons licence of this publication. For permission to reuse, please contact the rights holder.

two lateral teeth, and the bottom one the muscle that retracts the medial tooth. Pattern B1 leads to a mode of chewing in which teeth protract simultaneously, while pattern B2 leads to a pattern in which lateral teeth open while the medial tooth protracts.

Just as in our normal modes example above, the two modes of tooth operation are characterised by two individual patterns being in or out of phase with one another. And just as in the physics example, there are higher-level phenomena that are better explained by appealing to the mode of operation than to a component level description: if we want to explain how lobsters are capable of digesting a particularly tough meal, for example, we might want to explain how the meal stimulates the anterior gastric receptor that causes Type 2 CPG output.

Unlike in the physics example, we don't attach mathematical variable labels to the two patterns generated by the CPG and produce alternative equations. But in labelling the patterns, we have labelled the features relevant for certain causal dependencies. The relationship between these patterns and underlying neuronal variables also isn't made explicit, but it's implausible to think it can't be characterised mathematically; this is a very simple 10 neuron system, and the patterns in question are patterns of its output.

What we seem to have here is a new variable that takes two values. This variable plays a role in real causal dependencies governing lobster behaviour. Those causal dependencies only come into focus when we move to the pattern output level of description. The patterns themselves are robust under many changes to the underlying neuronal description. But yet the whole description here is perfectly amenable to reduction: the relationship between the neuronal variables and the pattern is straightforward. This seems to be a biological example of a weakly emergent level.

5 Conclusion

Our world admits of a sparsely levelled description of interesting, high-value levels. While not as sparse as the six levels proposed by Oppenheim and Putnam (1958), they are nonetheless much sparser than a notion of levels based on a simple change of scale. These levels are local, system-specific, and often reducible. They are also crucial for adequate description and explanation because they describe the real dependencies and relationships that we exploit in our science. It is an objective fact about the world that it admits of novel and robust levels of explanation. With the benefit of twenty-first-century hindsight, I think the fact is unsurprising—if we think hard about the world picture envisaged by our fictional philosopher, it proves obviously inadequate to explain Newtonian features of the world, let alone non-Newtonian ones. But the contrast with a world containing only cheap, plenitudinous levels is instructive; the difference is objective and empirically obvious.

I speak of levels of description and explanation. In List's terms,[14] this makes this an epistemic account of levels. I am happy with that label, but the issues here are not merely epistemic. Science is also in the business of telling us what kinds of object and property are out there in the world. These novel and robust levels of description and explanation describe and explain planets and viruses as well as particles and black holes. Science tells us that the world is structured in such a way as to contain non-fundamental objects and properties, and it is these objects and properties that populate levels worthy of the name.

References

Batterman, Robert W. 2002a. Asymptotics and the role of minimal models. *British Journal for the Philosophy of Science*, 53(1), 21.

Batterman, Robert W. 2002b. *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction and Emergence*. Oxford University Press.

Butterfield, Jeremy. 2011a. Emergence, reduction and supervenience: a varied landscape. *Foundations of Physics*, 41, 920–959.

Butterfield, Jeremy. 2011b. Less is different: emergence and reduction reconciled. *Foundations of Physics*, 41, 1065–1135.

[14] See Chapter 1, Section 2 in this volume.

Combes, Denis, Meyrand, Pierre, & Simmers, John. 2002. Motor pattern switching by an identified sensory neuron in the lobster stomatogastric system. In: K. Wiese (ed.), *The Crustacean Nervous System*, 582–590. K. Weise ed. Springer.

Dizadji-Bahmani, Foad, Frigg, Roman, & Hartmann, Stephan. 2010. Who's Afraid of Nagelian Reduction? *Erkenntnis*, 73(3), 393–412.

Franklin, Alexander. 2019. *Why Isn't There Only Physics?* Unpublished PhD Thesis, King's College London.

Franklin, Alexander. 2019. Universality reduced. *Philosophy of Science*, 86(5), 1295–1306.

Franklin, Alexander, & Knox, Eleanor. 2018. Emergence without limits: the case of phonons. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 64, 68–78.

Franklin, Alexander, & Robertson, Katie. 2021. *Emerging into the Rainforest: Emergence and Special Science Ontology*.

Kitano, Hiroaki. 2004. Biological robustness. *Nature Reviews Genetics*, 5(11), 826–837.

Knox, Eleanor. 2016. Abstraction and its limits: finding space for novel explanation. *Noûs*, 50(1), 41–60.

Knox, Eleanor. 2017. Novel explanation in the special sciences: lessons from physics. *Proceedings of the Aristotelian Society*, vol. 117, 123–140. Oxford University Press.

Ladyman, J., & Ross, D. 2007. *Every Thing Must Go: Metaphysics Naturalised*. Oxford University Press.

List, Christian. 2019. Levels: descriptive, explanatory, and ontological. *Noûs* 53(4), 852–883.

Mullan, Dermott J. 2022. *Physics of the Sun: A First Course*. CRC press.

Oppenheim, Paul, & Putnam, Hilary. 1958. Unity of science as a working hypothesis. In: Feigl, H. (ed), *Concepts, Theories, and the Mind-Body Problem*, 3–36. University of Minnesota Press, Minneapolis.

Robertson, Katie. 2022. In search of the holy grail: how to reduce the second law of thermodynamics. *British Journal for the Philosophy of Science*, 73(4), 987–1020.

Robertson, Katie. forthcoming. *Autonomy Generalised; or, Why Doesn't Physics Matter More? Ergo*.

Ross, Don. 2000. Rainforest realism: a Dennettian theory of Existence. In: Ross, D., Brook, A., and Thompson, D. (eds.), *Dennett's philosophy: A Comprehensive Assessment*, 147. MIT Press.

Wallace, David. 2021a. *Probability and Irreversibility in Modern Statistical Mechanics: Classical and Quantum*. arXiv preprint arXiv:2104.11223.

Wallace, David. 2021b. Stating structural realism: mathematics-first approaches to physics and metaphysics. *Philosophical Perspectives*, 36(1), 345–378.

Wilson, Alastair. Forthcoming. Metaphysical emergence as higher-level naturalness. In: Yates, David (ed.), *Rethinking Emergence*. Oxford University Press.

Wilson, Mark. 2006. *Wandering Significance: An Essay on Conceptual Behaviour*. Oxford University Press.

Woodward, James. 2016. The problem of variable choice. *Synthese*, 193(4), 1047–1072.

Woodward, James. 2021. Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance. *Synthese*, 198(1), 237–265.

11

Levels of Fundamentality in the Metaphysics of Physics

Karen Crowther

1 Introduction

Recently, work has begun to explore the variety of different notions of fundamentality in (philosophy of) physics, as well as the connections between these and the different ideas of fundamentality in metaphysics.[1] The present paper aims to continue this work. I point out that there is a plausible notion of relative fundamentality in physics whose 'hierarchy' tends to correlate with the one that is intuitively adopted by naturalised metaphysics in articulating its (standard, or roughly consensus) notion of relative fundamentality. Although not implied, we can—provisionally, and in the spirit of exploration—interpret this correlation as the two accounts of relative fundamentality essentially capturing the same relation. Doing so, however, we find some surprising results, both for the philosophy of physics as well as for metaphysics of physics.

The structure of the paper is as follows: Section 2 presents a short statement of what is meant by 'relative fundamentality' in metaphysics, remaining open as to whether it is to be defined in terms of *grounding* or *ontological dependence* relations (i.e., both types of relation are considered); Section 3 introduces the conception of relative fundamentality in physics which I argue most plausibly captures the asymmetric dependence that characterises such a relation; Section 4 explores the possibility of interpreting this conception of relative fundamentality in physics as capturing the same relation as the ideas of relative fundamentality in metaphysics; Section 5 discusses some implications of this interpretation, as well as two other possible interpretations of this idea of relative fundamentality, and Section 6 concludes.

2 Relative fundamentality in metaphysics

There are many different ways of characterising relative fundamentality in metaphysics, and many debates about the best way to do so. I take it, however, that there

Karen Crowther, *Levels of Fundamentality in the Metaphysics of Physics* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Karen Crowther 2024. DOI: 10.1093/oso/9780192862945.003.0012

are two main approaches: one utilises *grounding*, and the other *ontological dependence* relations (remaining neutral as to whether or not grounding is a species of ontological dependence relation, or vice versa). Grounding is supposed to be a relation between facts, while ontological dependence can hold between entities, properties, behaviours, events, or other relata.[2] So, fact *M* is more fundamental than fact *L*, if *M* grounds*L*. And entity/property/etc. *M* is more fundamental than entity/property/etc. *L*, if *L* is ontologically dependent upon *M*. If *L*—whether fact, or entity/property/etc.—is

grounded in, or ontologically dependent on *M*, then *M* is said to be *ontologically prior* to the less fundamental *L*. (Throughout this chapter, I use *M* to denote the *More* fundamental object (fact, or theory), and *L* to denote the *Less* fundamental object when considering relative fundamentality between two objects (facts, or theories)).

There are two key features of the grounding and ontological dependence relations that enable them to be used in defining relative fundamentality (RF). First is the fact that they are asymmetric relations: crucial to any conception of relative fundamentality is that it capture the idea that there is a *hierarchical structure* to reality (Bliss & Priest, 2018; Tahko, 2018). Second is the fact that they are not purely modal relations (Fine, 2012). Thus, I take these two minimal conditions to characterise the metaphysics conception of relative fundamentality:[3]

RF in metaphysics:

(i) *Asymmetry*: if fact/entity/event *M* is more fundamental than *L*, then it is not the case that *L* is more fundamental than *M*;

(ii) *Non-modality*: the relation must state something more than simply 'if *M*, then necessarily *L*'.

3 Relative fundamentality in physics

Physicists (and philosophers of physics) recognise a huge variety of features of physical theories as being indicative of a theory's (or entity's) status as either fundamental, or relatively fundamental.[4] Here, I take it that there is one minimal condition for relative fundamentality in physics: that it is a relation of *asymmetric dependence*. The relevant sense of asymmetric dependence is demonstrated in physics through *derivability* (note, by being evidence of asymmetric dependence, derivability is supposed to be *demonstrative* of fundamentality, rather than *constitutive* of fundamentality).

RF in physics: theory *M* is more fundamental than theory *L* if *L* is *derivable* from *M*, and *M* is not derivable from *L*. This demonstrates that *L asymmetrically depends* upon the physics described by *M*.

Here, *M* and *L* are physical theories, or parts of theories that are, were, or will be accepted by mainstream physics as approximately true. *M* and *L* are to be compared at a given time, in the form that each is accepted in at that time.

Derivability is a standard way of obtaining the physicists' sense of reduction as domain subsumption, i.e., the demonstration that all the successful parts of the reduced theory, *L*, can (approximately and appropriately) be obtained from the reducing theory, *M*. The idea of asymmetric dependence, or relative fundamentality, is that the reduced theory is shown, via reduction, to be embedded within the reducing theory (and hence that the physics described by the reduced theory depends on that of the reducing theory (Crowther, 2020)).

The claim of relative fundamentality as demonstrated by derivability underlies two of the most popular conceptions of relative fundamentality in physics: (i) a more fundamental theory as one applicable at *shorter length scales* (*M* describes 'smaller stuff' than *L* does); and (ii) a more fundamental theory as one with a *broader domain*, or increased generality (*M* describes 'more stuff' than *L* does). These two popular claims roughly correspond to what are typically distinguished as two different types of reduction in physics: (i) Reduction$_1$, and, (ii) Reduction$_2$. The distinction between two types of reduction was originally drawn by Nickles (1973).[5] *Reduction$_1$* is, roughly, the deduction of (corrected) parts of a higher-level theory from (parts of) a lower-level one, under some appropriate conditions, and is standardly exemplified by Nagel-Schaffner reduction (after Nagel, 1961; Schaffner, 1976). *Reduction$_2$* is, roughly, various inter-theory relations between (parts of) a new theory and its 'predecessor', under appropriate conditions, that serve heuristic and justificatory roles in theory succession.

Wimsatt (1976; 2006) elaborates on Nickles's distinction, and relabels it as one between *explanatory* and *successional* reduction. Explanatory reduction, according to Wimsatt, is an *inter-level* relation, relating 'levels of organisation' rather than theories (however, I will stick to speaking about theories in this chapter).[6] Its aim is to provide a *compositional*, *mechanistic*, and *causal* explanation of some large-scale phenomena in terms of shorter-length scale behaviours (Wimsatt, 2006, p. 4); e.g., explaining the behaviour of gases as clouds of colliding molecules, or the behaviour of genes in terms of the action of DNA (2006, p. 449). Wimsatt is clear that explanatory reduction is no longer best exemplified by Nagel-Schaffner reduction, but that it is richly complex and greatly diverse in its approaches, especially in biology. Successional reduction, according to Wimsatt, does relate theories, and is supposed to be *intra-level*: holding between newer and older theories, and/or more exact and more approximate theories, and/or more and less general theories that apply 'at the same compositional level'. But this sense of intra-level reduction is supposed to also relate theories that are not level-specific, such as in physics (Wimsatt, 2006, p. 450).

Here is how I use the terms in this chapter:

Reduction$_1$: 'explanatory reduction' holds between two theories formulated at different levels, e.g., different energy/length scales. *M* and *L* describe different degrees of freedom. *M* and *L* are related by *coarse-graining procedures*. *M* may be a 'constructive theory' or constitutive theory, i.e., describing particles or mechanisms supposedly underlying the physics described by *L*.

Reduction$_2$: 'successive reduction' holds between more general/less general, or more exact/more approximate theories at the same level. *M* and *L* may be overarching frameworks, or 'principle theories', not restricted to a certain level.[7] *L* is restricted in ways that are revealed and overcome by *M* (typically, the succeeding theory). *M* and *L* are related by the *weak field limit*.

These two notions tend to be exemplified together in physics, so that it is difficult to find 'pure' examples of either reduction$_1$ or reduction$_2$. A standard example of reduction$_1$ is the theory (or framework) of thermodynamics as less fundamental than statistical mechanics. Thermodynamics, which describes macroscopic physical quantities, is taken to be (in principle) derivable from statistical mechanics, which represents the underlying mechanistic theory of particles—the microscopic constituents of the thermodynamic system. Statistical mechanics is seen as providing an explanation of thermodynamic behaviour via the reduction, which utilises coarse-graining procedures.[8] Another example of reduction$_1$ is chiral perturbation theory as less fundamental than quantum chromodynamics (QCD). Chiral perturbation theory is an *effective field theory* constructed based on the symmetries of QCD which allows us to study the low-energy dynamics of QCD. The two theories describe different degrees of freedom: chiral perturbation theory is a theory of hadrons, which are supposed to be composite particles of quarks and gluons. The 'constituent' quarks and gluons are described by QCD at high-energy scales. A more general example is atomic theory as less fundamental than the standard model of quantum field theory, which describes the sub-atomic physics and provides the most fundamental description of matter.

In each case of reduction$_1$ the two theories apply at different length scales, or different levels, and they describe different degrees of freedom, with *M* being finer-grained, and *L* being coarser-grained. This conception of relative fundamentality has been heavily shaped by the development of the framework of effective field theory (EFT) and associated philosophy, especially in regard to discussion of emergence.[9] EFT is a toolbox for constructing macro theories valid at low energy scales from micro theories valid at high energy scales—e.g., the coarse-graining procedures. The resulting *effective* theories are not supposed to be fundamental, given their restricted domain of applicability. Nevertheless, they are extremely useful, being highly predictive and providing an understanding of the large-scale phenomena by being framed in the appropriate degrees of freedom for the length scales at which they apply (Georgi, 1989).

The high-energy *M* is more encompassing than *L*: in principle, it predicts everything that *L* does, but it also describes physics at energy scales where *L* does not apply.

Examples of reduction$_2$ include quantum electrodynamics as more fundamental than classical electrodynamics; special relativity as more fundamental than Newtonian mechanics; and quantum gravity as more fundamental than general relativity. These theories are *universal*: they are not supposed to be restricted to certain length scales. Yet, *M* replaces *L* as a more encompassing theory; *L* is restricted in some ways that are revealed and overcome by *M*, through the relation of reduction that connects them. This is typically demonstrated using the weak field limit (amongst other relations). For example, classical mechanics was seen as a universal theory of motion, but after the development of special relativity, Newtonian mechanics was shown to be invalid for velocities comparable to the speed of light. Special relativity is a more general, more encompassing theory, since it applies to everything that classical mechanics does, plus more. The dependence of classical mechanics on special relativity is shown through various relations that connect the theories, including the limiting relation of low velocities compared to the speed of light.

Reduction$_1$ is usually taken to be (potentially) ontologically interesting: the idea is that it is capable of providing a 'mechanism', a nice physical story, or a part-whole explanation of the higher-level phenomena. The higher-level phenomena described by *L* is considered 'real', and *L* is still retained as correct, as a 'special science'. Contrarily, reduction$_2$ is typically not thought to be ontologically significant; it is not thought to provide the same quality of explanation as 'explanatory reduction'. The older theory *L* is demoted as strictly false, but nevertheless may be considered 'approximately correct' in its restricted domain. The parts of *L* that are retained from the older theory are those that are shown to be compatible with (yielding approximately the same results as) the newer theory, *M*, via the reduction (Crowther, 2020).

I find this difference in attitudes to be ill-founded—it seems to be based purely on the fact that reduction$_1$ involves the idea of levels, related through coarse-graining procedures, while reduction$_2$ does not. Otherwise, there are many parallels between these two types of reduction. Both types utilise various inter-theory relationships aimed at establishing that *L* is in principle derivable from *M*, and thus that *M* subsumes the domain of *L*; i.e., *M* describes everything that *L* does, plus more. In both cases, *M* explains the phenomena described by *L* by being consistent with *L* in the relevant domain; i.e., *M* explains the success of *L* by yielding approximately the same results as *L* in the domain where *L* is known to be successful. This is true even in the case of inter-level 'explanatory' reduction$_1$. The 'lower-level' theories are more fundamental *not* because they provide a 'mechanism' or part-whole explanation, etc., but because *L* is (in principle, approximately) derivable from *M* and has broader domain. This establishes that *L* asymmetrically depends upon *M*.

An objection might claim that, in reduction$_2$, *L* cannot be of ontological significance because *L* is just a special case of *M*—the stuff described by *L* doesn't 'really exist' and *L* is in principle dispensable given *M*, e.g., classical systems 'are really just' quantum systems. But this same line of thought can be applied to the levels picture of reduction$_1$, where it echoes *naïve reductionism*. According to naïve reductionism, any emergent physics, or phenomena described by special sciences, doesn't 'really exist', and the higher-level *L* theories are in principle dispensable. Not many people support such a naïve reductionist attitude today, and to be clear, neither do I—but nor do I accept the analogous position in the non-levels case of reduction$_2$. In both cases, *L* theories continue to be used—not just because of practical necessity, but because they provide a useful description of physics that is appropriate to the domains in which we most often find ourselves. The less fundamental theories impart an understanding of the phenomena, and the more fundamental ones (in the absence of the connecting relations) do not (Crowther, 2015).

A more general objection might be that the idea of derivability between theories is not always ontologically significant, since it could instead be merely mathematical relations connecting the theories. This is, however, not true in the case of the physicists' sense of reduction in general. It is a requirement upon any new theory of physics that it appropriately connect with its predecessor (including a 'higher-level' theory) via this idea of reduction, which necessarily involves some reasonable physical interpretation. It is generally seen as a problem if no physical sense can be made of the relations connecting the theories (even if, as typical, these relations involve approximations, and *L* may be seen as in some sense an approximation of *M*). It should also be emphasised that these sorts of reductions are not typically just a single mathematical relation such as a limiting relation linking the theories, but instead involve establishing various types of connections ('correspondence relations', which are not all mathematical in nature) between the theories (Crowther, 2020). Finally, this chapter is concerned with the metaphysics of physics, and involves asking what our physical theories—in the form in which they are currently accepted—could tell us about the ontology of the world (objects and/or relations), so the spirit is a realist interpretation of our theories and the relations between them.

So, I argue that both reduction$_1$ and reduction$_2$ are means of establishing that *L* is derivable from *M*, and thus that *M* is more fundamental than *L*. In other words, the derivability demonstrates that *L* is asymmetrically dependent upon *M*. But what kind of dependence does derivability establish? Most straightforwardly, it captures modal dependence (but we will see later that there is also a notion of natural dependence and ontological dependence). Theory *L* is less fundamental than theory *M* if *L* depends upon *M*, but *M* is not necessary for *L*. If *M* holds in the world, then necessarily so does *L*. *L* is derivative: it holds in the world because *M* does. But it is possible that (or we can imagine a world where) we have *L* without *M* (i.e., there may be other reasons why *L* obtains, other than *M* obtaining).

This is easier to understand by looking at the examples. Newtonian gravity is less fundamental than general relativity (GR). If GR holds in our world, then necessarily so does Newtonian gravity, since it is the weak field limit of this theory, it is 'contained' within GR. Newtonian gravity is derivative, it holds in the world *because* GR does. However, it is possible that there be a world where Newtonian gravity holds, but GR does not—Newtonian gravity might, e.g., be the weak field limit of a *different* more general gravitational theory (other than GR), or it may be that the world is *only* described by Newtonian gravity and there is no more fundamental theory of gravity. This is true also with 'inter-level' reduction. Atomic theory is less fundamental than the standard model of quantum field theory (QFT), and is derivative from it. If the standard model is true in our world, then necessarily so is atomic theory, since atomic theory is a low-energy limit of the standard model. But we can imagine a world where atomic theory is true, with a different more fundamental theory 'underlying' it. Another example is the current situation with respect to the theory of quantum gravity, which is supposed to be more fundamental than GR according to both senses of relative fundamentality described above. There are several different possible theories of quantum gravity from which GR could be derived, and in this sense GR is *multiply realisable*.[10] This is in spite of the fact that we believe there is only one correct theory of quantum gravity that holds in our world, and that it is by virtue of this that GR holds in our world.

4 Can we understand these relations as capturing the same idea?

If we understand relative fundamentality in physics as described above, we can speak of a hierarchy of more and less fundamental theories, which needn't be associated with 'levels' in the sense of theories applicable at different energy or length scales. Instead, the levels are *levels of fundamentality*, distinguishing the derivative from its 'basis'. Next, notice that the hierarchy of facts described by the hierarchy of physical theories tends to correlate with the chain of grounding relations, where the grounded fact (to be explained) is less fundamental than the fact grounding it (the explanation). Some examples:

- facts about the existence and behaviour of atoms, as described by atomic theory, are grounded in facts about sub-atomic physics, as described by QFT;
- facts about thermodynamic systems are grounded in statistical-mechanical facts; facts about electric current, as described by electrodynamics, are grounded in facts about quantum fields, as described by QFT;

- facts about the behaviour of systems at familiar velocities, as described by classical mechanics, are grounded in relativistic facts, as described by special relativity;
- facts about classical systems, as described by classical mechanics, are grounded in facts about quantum systems, as described by quantum mechanics.

Plausibly, then, there is a sense in which the hierarchy of physical theories captures grounding relations (however uneven or 'wobbly' the levels may be). This type of grounding relation is what Fine (2012) calls 'natural grounding'. This form of grounding is not as 'strict' as what Fine calls 'metaphysical grounding', which holds when the 'the explanans or explanantia are *constitutive of* the explanandum, or that the explanandum's holding *consists in nothing more than* the obtaining of the explanans or explanantia' (p. 37). The idea is that metaphysical grounding leaves no 'explanatory gap' between the grounded fact and the fact doing the grounding. In the case of natural grounding, however, there may be a gap, such that the fact doing the grounding is not *constitutive of* the grounded fact. To see the difference, we can compare the examples of natural necessity in the list above with Fine's (2012, p. 36) example of metaphysical necessity, as 'the fact that the ball is red and round is grounded in the fact that it is red and the fact that it is round'. Clearly, facts about atoms are further away from facts about sub-atomic particles than the fact of 'being red and round' is to 'being red and being round'.

Grounding relations are not purely modal claims, but express an explanatory or determinative connection between the two facts (Fine, 2012). Ontological, or metaphysical, grounding is the strongest form of connection, and is of special interest to metaphysics, while natural grounding is weaker, and of special interest to science (p. 36). One may thus object that the grounding relations described in the list above, correlating with the hierarchy of relative fundamentality in physics, reflect merely natural dependence rather than ontological dependence, and that only the latter is of interest to metaphysics. However, recall that here we are not doing pure metaphysics—we are doing *metaphysics of physics*, exploring what our theories of physics, on a realist interpretation, tell us about the entities and relations that exist in the world.

In this spirit, I claim that the entities described by each level in the hierarchy of physical theories tend to correlate with the chain of ontological dependence:

- the existence and behaviour of atoms, as described by atomic theory, ontologically depends upon sub-atomic physics, as described by the standard model of QFT;
- thermodynamic systems ontologically depend upon statistical mechanical systems;
- an electric current, as described by electrodynamics, ontologically depends on quantum fields, as described by quantum electrodynamics;

- systems at familiar velocities, as described by classical mechanics, ontologically depend on relativistic systems, as described by relativity;
- classical systems, as described by classical mechanics, ontologically depend upon quantum systems, as described by quantum mechanics.

The first two examples on the list reflect the relation of reduction$_1$, and may be easier to intuitively accept than the others. Of course, atoms are different things than sub-atomic particles or quantum fields, and we naturally think, too, that an atom is something different than the sub-atomic particles or quantum fields that 'underlie it' or 'compose it'—even though the atom's existence and behaviour is reducible to that of the quantum fields, and so, in a sense, the atom is *nothing other than* these fields on a different level of description. Contrarily, it may not feel so natural to think of a given non-relativistic object (i.e., an object not under the conditions where relativity is necessary in order to describe it) as being a different thing than its relativistic self—we can speak of both, but it intuitively seems more like two different descriptions of the one object, rather than two different objects. However, the only relevant difference between the atomic/sub-atomic example and the non-relativistic/relativistic example is that the first involves the idea of levels related through coarse-graining procedures and the other does not. Without an independent reason for thinking this relation is special in imparting ontological significance,[11] we may consider all the examples on the list as on par in capturing ontological priority. Hence, we may plausibly claim that the relation of relative fundamentality in physics correlates with that of metaphysics.

5 Implications and interpretations

In this section, I discuss different ways in which we might interpret the metaphysics of Sections 3 and 4: what does this striking conception of relative fundamentality tell us about what exists? The main implication seems to be that we can have levels of fundamentality (ontological priority) associated with hierarchies of theories related through the physicists' sense of reduction. Importantly, these levels are not universally defined: they are not associated with particular energy or length scales (i.e., the levels are not 'micro' versus 'macro' theories), but hold between any two theories where one theory is supposed to subsume the domain of the other and to be responsible for the success of the other, as demonstrated by *L* being derivable from *M*. Accepting both the *L* and *M* theories (so long as the *L* theories continue to be used in physics), we apparently get a 'rich ontology', which seems to admit both atoms plus sub-atomic particles, classical forces plus quantum fields, relativistic masses and non-relativistic masses, etc., as well as the reduction or grounding relations between them. But this is not the only interpretation.

Following Le Bihan (2018), we can recognise three different possibilities: (1) the *derivative view*; (2) the *eliminativist view*; and (3) the *reductionist view*.

The *derivative view* is the view just described, which implies the existence of levels of reality (again, as understood as levels of fundamentality). On this view, we accept the existence of everything at each level, we accept the existence of constitutive/building relations between entities at each level, and we accept relative fundamentality of the entities described by *M* compared to *L* (Le Bihan, 2018, §4). The entities of the *L* theories are real, but they are *derivatively real*: they are grounded in, or built from, the more fundamental ontology described by the corresponding *M* theories in each case. The main objection to this view is the ontological cost: we have a very rich ontology, with entities at all levels, as well as connecting relations that exist. Some other concerns with this view are expressed in Le Bihan (2018, pp. 82–83),

> The notion of ontological level is not very clear, at least not as much as the notion of descriptive level. What does it mean that behind levels of description (think for instance of the biological level or the chemical level) lie 'ontological levels'? One could argue that levels come for free and should not be interpreted too seriously. However, if ontological levels come for free, then these merely are levels of description: the notion of ontological level has no counterpart obtaining in the world. The derivative view thereby collapses into eliminativism.
>
> But perhaps one may argue that this is not a genuine problem. After all, maybe the ontological cost is well motivated insofar as it offers an adequate characterization of the delicate situation we face in contemporary physics. Nonetheless, if it is possible to come up with a view that does not entail the existence of levels of reality and has the same power of explanation, it should be preferred over the derivative space view.

So, the rich ontology may be seen as a high cost of the view, and we may be concerned that interpreting our theories as capturing genuine ontological levels rather than just levels of description is an ontological burden we needn't bear if avoidable. Nevertheless, we are here interested in the idea of relative fundamentality in physics, and this derivative view is the option to choose if we want to talk about relative fundamentality and want to take physics seriously—the two alternative views do not feature relative fundamentality.

The next alternative is the *eliminativist view*, which holds that the derivative entities of *L* are not real, and that only the ontology described by the fundamental *M* theories exists. Thus, there is no relative fundamentality, or levels of reality. This view has a minimalist ontology, but has the consequence that much of what we take to be true in physics is literally false: there are no atoms, no thermodynamic systems, no classical or non-relativistic systems at all. This becomes more disturbing with the recognition that physicists do not consider our current theories as fundamental,[12] and so according to this view we'd have reason to believe that *nothing* currently described by physics exists. Only the entities described by the (absolute) fundamental theory, *F*, exist: those of *M* and *L* do not. The idea is that *M* and *L* describe concepts rather than entities, and that these do not map neatly onto the entities of *F*.

This view departs from the naïve realism of the derivative view, and so represents a more sophisticated option. In other words, it requires more work. If we are to adopt this view, we need to address questions such as how it is that physics is so successful if the entities it describes do not exist, and how the concepts described by physical theories relate to the minimal ontology. If these questions are not seriously addressed, then adopting this view no longer counts as metaphysics of physics, but falls into pure metaphysics.

The third option is the *reductionist view*, according to which we accept the existence of derivative entities, but reject the existence of substantive constitutive/building relations between the 'derivative' and 'fundamental' entities—i.e., there are no levels of reality, no notion of relative fundamentality, and thus no genuine distinction between fundamental and derivative entities (understanding fundamentality as ontological priority). Instead, the derivative entities are (in a non-spatiotemporal sense) 'within' the fundamental structures, and thus the view is consistent with a *reductionist ontology* (Le Bihan, 2018, p. 84). This is the view favoured by Le Bihan in regards to the relationship between the spatiotemporal ontology of general relativity (as theory *L*) and the non-spatiotemporal ontology of quantum gravity (as theory *M*). The idea is that there is a weak relation of composition between the ontology of *M* and that of *L*, but this does not establish relative fundamentality of the *M* entities compared to the *L* ones. This weak relation of composition is supposed to be a non-spatiotemporal form of mereology, e.g., the relation of *logical mereology* of Paul (2002).

This reductionist view is the option to take if we don't want to talk about relative fundamentality, but still want to take physics seriously. Adopting this view does require some work, however, e.g., in elaborating the composition relation and explaining how it relates to the inter-theory relations of physics.

6 Conclusion

I have argued that one way of understanding relative fundamentality in physics is as holding between two theories, *M* and *L*, where the more fundamental theory *M* has a broader domain than *L*, and *L* is derivable from *M*. This establishes that *L* asymmetrically depends upon the physics of *M*. Such a relation of relative fundamentality needn't be associated with micro and macro theories, but can also hold between theories that are not restricted to certain length scales—e.g., *M* may be a more general theory, more exact, or finer-grained than *L*. This chain of relative fundamentality can plausibly be seen as correlating with an idea of relative fundamentality in metaphysics: facts about the physics described by *L* are grounded in facts described by *M*. This can be understood plainly as expressing natural dependence, but I further speculated, based on the potential fruitfulness of doing so, that we can understand this as also capturing ontological dependence—i.e., that the existence and behaviour of the entities of *L* ontologically depend on those of *M*.

Viewing the physics of a less fundamental theory as ontologically dependent on that described by a more fundamental theory leads to some startling consequences for ontology, however. For instance, that classical systems ontologically depend upon quantum systems just as thermodynamic systems ontologically depend upon statistical mechanical systems, or as atoms depend on sub-atomic physics. While this seemed to commit us to a rich ontology including derivative as well as fundamental objects plus the dependence relations, I argued, following Le Bihan (2018), that there are two other possible interpretations. We are only committed to the rich ontology if we want to retain a notion of relative fundamentality.

Notes

[1] Specifically: Morganti (2020a, b); an international workshop organised by Fabrice Correia, Claudio Calosi, and Benjamin Neeser, Geneva, 2018; symposia at the BSPS conference 2018, Oxford, and the PSA 2018, Pittsburgh. Also related are the dedicated attempts within naturalised metaphysics to genuinely explore the meaning, and implications of, particular physical theories for the metaphysics of fundamentality, or to apply the 'tools' of metaphysics to better understand the ideas of fundamentality suggested by particular theories of physics, e.g., Le Bihan (2018); Le Bihan & Read (2018); McKenzie (2011, 2017).

[2] Although there are more differences between these two types of dependence relation than merely their differing relata. See, e.g., Calosi (2020); Kovacs (2018, 2019).

[3] This is following a suggestion by Andreas Hüttemann.

[4] See, e.g., Cao (2003); Crowther (2019); Morganti (2020b).

[5] This distinction has also been referred to as *explanatory* versus *successional* reduction (Wimsatt, 1976, 2006), *synchronic* versus *diachronic* reduction (Dizadji-Bahmani et al., 2010; Rosenberg, 2006; van Riel & Van Gulick, 2016), and *vertical* versus *horizontal* reduction (Robertson & Wilson, forthcoming).

[6] Wimsatt (1976, p. 680) conceives of levels of organisation as 'primarily characterized as local maxima of regularity and predictability in the phase space of different models of organization of matter'.

[7] Famously, the distinction between constructive theories and principle theories is from Einstein's 1919 article 'What is the Theory of Relativity' in *The Times* (Einstein, 1954).

[8] This particular example has been heavily debated as to whether or not it represents Nagelian reduction, but for our purposes of illustrating reduction$_1$, which need not be strict Nagelian reduction, the example is apt.

[9] See, e.g., Bain (2013); Castellani (2002); Crowther (2015).

[10] Cf. Crowther (2020); Jaksland (2019).

[11] Those who already find this relation to be important in understanding ontological emergence may have an argument here, however.

[12] See Crowther (2019).

References

Bain, J. (2013). Effective field theories. In B. Batterman (Ed.), *The Oxford Handbook of Philosophy of Physics* (pp. 224–254). New York: Oxford University Press.

Bliss, R., & Priest, G. (2018). The geography of fundamentality: an overview. In R. Bliss & G. Priest (Eds.), *Reality and its Structure: Essays in Fundamentality* (pp. 1–35). Oxford: Oxford University Press.

Calosi, C. (2020). Priority monism, dependence and fundamentality. *Philosophical Studies*, *177* (1), 1–20.

Cao, T. Y. (2003). Appendix: ontological relativity and fundamentality—is QFT the fundamental theory? *Synthese*, *136* (1), 25–30.

Castellani, E. (2002). Reductionism, emergence, and effective field theories. *Studies in History and Philosophy of Modern Physics*, *33* (2), 251–267.

Crowther, K. (2015). Decoupling emergence and reduction in physics. *European Journal for Philosophy of Science*, *5* (3), 419–445.

Crowther, K. (2019). When do we stop digging? Conditions on a fundamental theory of physics. In A. Aguirre, B. Foster, & Z. Merali (Eds.), *What is 'Fundamental'?* (pp. 123–133). Springer.

Crowther, K. (2020). What is the point of reduction in science? *Erkenntnis*, *85* (6), 1437–1460.

Dizadji-Bahmani, F., Frigg, R., & Hartmann, S. (2010). Who's afraid of Nagelian reduction? *Erkenntnis*, *73* (3), 393–412.

Einstein, A. (1954). What is the theory of relativity? In *Ideas and Opinions* (pp. 227–232). New York: Bonanza.

Fine, K. (2012). Guide to ground. In F. Correia & B. Schnieder (Eds.), *Metaphysical Grounding* (pp. 37–80). Cambridge University Press.

Georgi, H. (1989). Effective quantum field theories. In P. C. W. Davies (Ed.), *The New Physics* (pp. 446–457). Cambridge: Cambridge University Press.

Jaksland, R. (2019). The multiple realizability of general relativity in quantum gravity. *Synthese*, *199* (S2), 441–467.

Kovacs, D. M. (2018). The deflationary theory of ontological dependence. *Philosophical Quarterly*, *68* (272), 481–502.

Kovacs, D. M. (2019). Ricki Bliss and Graham Priest (eds.): reality and its structure: essays in fundamentality. *Notre Dame Philosophical Reviews*, *17*.

Le Bihan, B. (2018). Space emergence in contemporary physics: why we do not need fundamentality, layers of reality and emergence. *Disputatio*, *10* (49), 71–95.

Le Bihan, B., & Read, J. (2018). Duality and ontology. *Philosophy Compass*, *13* (12), e12555.

McKenzie, K. (2011). Arguing against fundamentality. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *42* (4), 244–255.

McKenzie, K. (2017). Relativities of fundamentality. *Studies in History and Philosophy of Science Part B*: *Studies in History and Philosophy of Modern Physics*, *59* (Supplement C), 89–99. Dualities in Physics.

Morganti, M. (2020a). Fundamentality in metaphysics and the philosophy of physics. part i: Metaphysics. *Philosophy Compass*, *15* (7), e12690.

Morganti, M. (2020b). Fundamentality in metaphysics and the philosophy of physics. part ii: the philosophy of physics. *Philosophy Compass*, *15* (10), e12703.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt.

Nickles, T. (1973). Two concepts of intertheoretic reduction. *Journal of Philosophy*, *70* (7), 181–201.

Paul, L. A. (2002). Logical parts. *Noûs*, *36* (4), 578–596.

Robertson, K., & Wilson, A. (forthcoming). Theoretical relicts: progress, reduction, and autonomy. *British Journal for the Philosophy of Science*.

Rosenberg, A. (2006). *Darwinian reductionism: Or, How to Stop Worrying and Love Molecular Biology*. Chicago: University of Chicago Press.

Schaffner, K. (1976). Reductionism in biology: prospects and problems. In Cohen, R. (Ed.), *PSA 1974* (pp. 613–632). D. Reidel Publishing Company.

Tahko, T. (2018). Fundamentality. In E. Zalta (Ed.), *Stanford Encyclopedia of Philosophy*. Stanford University.

van Riel, R., & Van Gulick, R. (2016). Scientific reduction. *Stanford Encyclopedia of Philosophy*. Stanford University.

Wimsatt, W. C. (1976). Reductive explanation: a functional account. *PSA 1974*, 671–710.

Wimsatt, W. C. (2006). Reductionism and its heuristics: making methodological reductionism honest. *Synthese*, *151* (3), 445–475.

12

No Grounds for Effective Theories

Kerry McKenzie

1 Introduction

It is by now a familiar idea that there has been an 'explosion' of work in 'stratified metaphysics': that is, in metaphysical projects aimed at articulating the idea that reality comes parceled into a structure of 'layers' or 'levels'. And largely constitutive of these efforts has been the work that has gone into understanding the notion of 'ground'. Ground has after all emerged as the preferred candidate for the 'level connector' in metaphysics given the well-catalogued failures of purely modal conceptions, such as supervenience.[1] As such, there has been something of a 'grounding revolution' afoot in contemporary metaphysics for over a decade now (Schaffer 2009; Kovacs 2017, p. 2927). Part of the motivation for so much work on grounding has been the recognition that many canonical metaphysical questions seem at their heart to be about ontological priority, and hence about what is more fundamental than what. But a second and distinct motivation is that the world described by the sciences seems to come stratified into levels. As such, any metaphysics that aims to be adequate to the sciences ought to capture this fact, and this is seen as more work for ground to do. As Schaffer puts it, 'grounding is a notion that is extremely natural in the sciences, in considering the relation between levels. One need not be versed in an arcane metaphysics to think that the chemical depends on the physical' (Schaffer 2016, section 4.4). There is thus motivation to study ground internal to metaphysics, but apparently also coming from the sciences.[2]

The purpose of this chapter is to pressurize this idea that the grounding relation can be regarded as the 'level-connector', insofar as the levels concern something described by the sciences.[3] My strategy will not be to problematize the idea that the sciences can faithfully be represented as describing a world that admits

[1] As Sider puts it, 'Metaphysics has always needed a "level-connector". One doesn't get far in metaphysics without some sort of distinction between fundamental and non-fundamental facts, or between more and less fundamental facts … So there's a niche for a metaphysical but nonmodal conception of the connection between levels. That niche has been filled by ground' (Sider 2020, pp. 747–778). By now the classic statement of this is probably Schaffer (2009).

[2] See McKenzie (2022, Ch. 3) for further discussion of the two sources of inspiration for grounding.

[3] I will remain neutral here on whether it fares better in illuminating canonical questions in metaphysics: though see McKenzie (2022, Ch. 3) for other reasons to think the relations involved in each context must be different.

Kerry McKenzie, *No Grounds for Effective Theories* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson,
Oxford University Press. © Kerry McKenzie 2024. DOI: 10.1093/oso/9780192862945.003.0013

of some kind of levels structure, although this is a path that has been taken by others. Rather, I will argue that while science does describe what we may regard as 'levels in nature', the relation of grounding cannot be taken to be that which links them together. The reason is that some core principles standardly taken to characterize grounding are incompatible with the relationship that exists between more and less fundamental *quantum field theories*. While such theories prima facie describe only a part of the hierarchy of levels—having nothing direct to say about the relation between, say, proteins and living cells—this is nevertheless important for several reasons. The first is that if the argument is correct, it implies that grounding is not the level-connector *simpliciter* since it does not have the generality that has been claimed for it. But a further and more fundamental reason is that it is arguably quantum field theory, and in particular the notion of 'effective field theories' (EFTs) that it sanctions, that supplies an explanation of *why* there are levels in physics at all—an assumption whose truth is by no means obvious. Since it is the effective field framework that gives a systematic physical explanation of the very existence of a levels structure, a metaphysics that aims to articulate the general nature of the 'level connector' has particular reason to be adequate to it.

The structure of my argument is as follows. In Section 2, I outline some of the reasons why the idea of 'levels in nature' is in certain respects a puzzling one. In Section 3, I outline why it is the emergence of the effective field theory concept that offers an explanation of why there are levels in physics. In Section 4, I outline some of the properties standardly attributed to grounding and, by drawing on some classic literature within philosophy of science, argue that they are not compatible with the relation that connects successive effective field theories. Section 5 is a brief conclusion. Throughout, I will take a 'level' in physics to be the sort of thing that can be described by a theory, and to comprise 'a domain with its own set of entities, structures, and laws' (Rivat and Grinbaum 2020, p. 90). I will take grounding to be a worldly relation, and I will assume that relating 'levels' essentially involves relating the relevant laws.[4]

Potochnik (2017, Ch. 6) offers an extended critique of the notion of levels. Ladyman and Ross go so far as to say that 'contemporary science gives no interesting content' to the metaphor of 'levels of reality' (Ladyman and Ross 2007, p. 54) and even describe it as 'profoundly unscientific' (p. 4). I will not follow them here: see Craver (2015) and Havstad (forthcoming) for what I take to be some compelling reasons why.

[4] More on this in Section 5 below. I note for now that it is by this point quite standard to talk of non-fundamental laws as being 'grounded' in the more fundamental laws: see, e.g., Bhogal (2017).

2 'Levels of nature' and the puzzle of quasi-autonomy

While some have used the concept of effective field theories to challenge the existence of a fundamental level (Cao and Schweber 1993), it will be easier to introduce the issue presuming that there is such a level and that QFT provides a description of it. What we are assuming, then, is that the basic principles of QFT furnish a theoretical description of whatever fundamental laws, properties, and objects the world contains. It is of course conceded at this point that we do not know what this specific quantum field theory is: even our most fundamental current theory, the Standard Model of particle physics, is regarded as merely an *effective* theory (see below) of an unknown but truly fundamental theory of quantum gravity.[5] We do believe, however, that we nevertheless know plenty of physics; indeed it is this physics that is largely guiding our investigations towards the elusive fundamental theory.

An immediate consequence of this situation is that the fundamental theory, while taken to *determine* the rest of physics, is nevertheless in some sense *distinct* from it. This is of course part of why it is we talk of different 'levels' in the first place: it is because the phenomena of the world seem to be largely confinable into different regimes—regimes that roughly correlate with size or spatial 'scale'. That is, some features manifest at cosmological scales; others at more mundane macroscopic scales; and other features seem to be manifest at microscopic scales alone. And while it is a basic presumption of the 'levels hierarchy' that these are in some way related, it is nevertheless implicit in scientific practice that each can be studied relatively independently of the others. Indeed, it is even conceded by practitioners that one usually does better by studying a certain level independently of the others. As one field theorist puts it:

> It is a basic fact of life that Nature comes to us in many scales. Galaxies, planets, aardvarks, molecules, atoms and nuclei are very different sizes, and are held together with very different binding energies. Happily enough, it is another fact of life that we don't need to understand what is going on at all scales at once in order to figure out how Nature works at a particular scale. Like good musicians, good physicists know which scales are relevant for which compositions. (Burgess 2007, p. 330)

Given that human beings are not omniscient with respect to the subject matter of physics, were it not for this effective separation of scales then it seems 'physics as we know it would be impossible' (Van Kolck, Abu-Raddad, and Cardamone 2002, p. 196). But while this may be a 'basic fact of life', there is in fact nothing obvious

[5] For a highly contemporary expression of this viewpoint, see Weinberg (2021).

about why nature should be organizable in this way. For this levels picture assumes a certain *autonomy* between levels, such that upper levels may be meaningfully theorized largely without consideration of the more fundamental. However, there is nothing obvious in general about why a domain should be able to be successfully theorized in ignorance of the fundamental principles that govern it. And in the case of QFT it is in fact especially mysterious as to why it should be that less fundamental phenomena that lie within its scope should be describable independently of fundamental goings-on. QFT is after all a 'local' theory, meaning it studies interactions between fields at a point. But quantum uncertainty then requires that processes of arbitrarily high momenta must be included in the calculations of the probable outcomes of such interactions, whatever the fields involved (since the smaller the spatial domain considered, the larger the range of relevant momenta). Since it is in the short distance—equivalently, high-energy or high-momenta—domain that we take fundamental processes to hold sway, it seems that low-energy QFTs in principle *cannot* be understood without a grip on the fundamental theory.

It is therefore somewhat ironic that it is within QFT that we arguably find the most 'systematic and controlled' method for deriving the relations of relative autonomy between theories that allow us to stake out different 'levels' (Hartmann 2001, p. 268). It was the development of the concept of the *effective field theory* that paved the way for a nuanced understanding here. As Hartman puts it:

> It is not an easy task to make more precise what it means exactly that different levels of organisation are autonomous. However, within the programme of EFTs, the notion of quasi-autonomy can be given a precise meaning and the relation of one level of organisation to a deeper level can be studied. (Hartmann 2001, p. 269)

My contention in this paper is that this relation between 'levels of organisation' is not the relation of grounding. To show why, I will first outline the basic features of how, at least in straightforward cases, the 'EFT programme' delivers a scientifically precise notion of a 'levels structure' and thus a clear sense to relative fundamentality—something that, as I hope to make clear, represents a real scientific and philosophical achievement.[6] Following that, I will defend the claim that the 'level connectors' used to map the levels structure so obtained do not have the features that are standardly taken to be essential to grounding.

[6] There are conditions under which some of the following statements do not hold. For example, turbulent systems do not exhibit the separation of scales required to get the EFT machinery running. But in such systems we might be hesitant to talk about levels at all. In any case, my argument only needs to apply to a part of the 'levels hierarchy' for it to have bite.

3 The emergence of the 'effective field theory' concept

Our understanding of how the framework of QFT delivers back the notion of a 'hierarchy of levels' came via the attempt to understand the process of 'renormalization' that necessarily accompanies the extraction of predictions from interacting QFTs. As such, I will say just a little about this process and why the need for it arises.[7] To begin at the beginning: a QFT is a theory of the interactions of *fields* which respects the *principles of relativity* and the *principles of quantum mechanics*. Among the latter principles is the principle of *unitarity*, which requires that the probabilities of experimental outcomes, as computed by the Born rule, always sum to one. The interactions of these fields are typically given by a *Lagrangian*, which may be regarded here as interchangeable with a 'law of nature'—something that describes that way that the entities in the domain in question evolve and interact.[8] In many ways the simplest example of an interaction Lagrangian in QFT is so-called ϕ^4 theory, which describes the interactions of a scalar field $\phi(x)$ with itself via a quartic self-interaction term:

$$L = \frac{1}{2}\left(\partial_{\mu}\phi\right)^{2} - \frac{1}{2}m^{2}\phi^{2} - g\phi^{4} \tag{1}$$

The g in this theory is the 'coupling constant' and represents the strength of the field's self-interactions. Field theorists will attempt to extract empirical predictions from this theory by calculating the scattering matrix (S-matrix) of the theory—a matrix whose elements give the *probabilities* corresponding for encounters between the particles described by the theory. It turns out that if g is large then there is no general method for extracting these amplitudes. However, assuming that the interaction coupling g is small—that is, < 1 in suitable units—the probabilities for obtaining a particular output state (such as the production of a certain particle) given a certain input state (paradigmatically some particles colliding with each other) may then be expressed as a series arranged in powers of g:

$$Prob(output,input) = \sum_{n=0}^{\infty} g^{n} \int_{E=0}^{E=\infty} F_{n}(E)dE \tag{2}$$

Here we don't need to worry too much about the functions F_n beyond that (i) they are functions of the energy and (ii) they are integrated over, with the range of integration unbounded from above. It is the fact that Minkowski space is

[7] For a lucid presentation of the ideas behind renormalization, see Williams (2021).

[8] One gets from the Lagrangian to the familiar 'laws of temporal evolution' via the Euler-Lagrange equations. Note also that, while what one uses to compute is the law statement, in accordance with standard usage I'll refer to both law statements and the pattern in properties that they refer to as the 'laws'.

continuous, and thus resolves into regions that are arbitrarily small, that accounts for why the range of integration is unbounded from above; and since (to repeat) this is a local theory it seems all of this infinite range must be taken into account in computing the amplitude. Unfortunately however—although not particularly surprisingly—what one finds is that these integrals generically *diverge*, leading to an egregious violation of unitarity and to mathematical nonsense. This 'ultraviolet catastrophe' for a long time made it look like something was remiss in the very foundations of the theory. Rather than give up on the QFT framework, however, what was developed was a means of 'taming' these infinities such that sensible predictions could be extracted from the theory. This arduous and initially perplexing process is known as the process of 'renormalization'.

The initial idea behind renormalization was that these divergences were a manifestation of the Lagrangian used to compute amplitudes being somehow incorrect, or at least incomplete. As such, additional terms could be added to it to see if they helped to bring things under control. What was found was that in a certain class of theories—the 'renormalizable' theories—*finitely many* terms could be added to make the divergences disappear. Since as a general rule these terms have to have the same form as the original terms to cancel the infinities appropriately, renormalization in essence amounts to making a change in the value of the theory's coupling parameters (such as *g*), from a finite to an infinite value. Implausible as it may seem, if done correctly this succeeds in restoring finitude to the amplitudes and to the possibility of it making empirical predictions. Indeed, renormalized QFTs such as QED have arguably resulted in the most accurate predictive successes of all time.

For all that, and while not entirely ad hoc from a physical point of view, the renormalization process seems to amount to a rather dark art mathematically speaking.[9] As such, it was for some time a source of some embarrassment among the physics community: Richard Feynman famously spoke of it as 'sweeping infinities under the rug'. But even aside from one's views about the legitimacy of the technique from a conceptual point of view, it is actually very surprising that such a procedure can even be made to work at the formal level. It is, after all, not unreasonable to presume a priori that the most energetic processes involved in an interaction make the most important contributions to it, and hence to the associated amplitude. As such, it is very counter-intuitive that their contribution could be modelled in such a simple way as to modify the constants in it.[10]

The work of Kenneth Wilson is largely taken to have finally explicated why it is that renormalization works. What Wilson realized was that if we are to understand what is going on in the process of renormalization then we have to explain why

[9] It is not entirely ad hoc because one can argue that measured charge of an electron partially results from the 'screening' effects of virtual particles. As one moves closer to the electron these effects decrease, and one is left with an unmeasurable 'bare charge' whose value, for all we know, could be infinite.

[10] Peskin and Schroeder (1995, p. 393).

the high-energy contributions of the theory can be modelled as they are in that process. And in order to explain that, we need to understand the effects that the short-distance degrees of freedom have on the interaction amplitudes and other observable quantities. These observable quantities are ultimately determined by the theory's 'S-matrix', the elements of which may in turn be calculated from the *path integral*

$$Z = \int D\phi \exp - i \int L\left(\phi,\ \partial_x \phi\right) d^4 x \tag{3}$$

where L is the Lagrangian describing the interaction of the fields φ. The quantity $L(\phi,\partial_x\phi)d^4x$ is known as the *action*, and the $D\phi$ indicates that we are integrating over all configurations of the field that have as their boundary conditions the given input and output states. What we are interested in is determining the effect that high-energy processes have on the low-energy interactions of this field. To see that, let us now divide the field into high and low energy components ϕ_H and ϕ_L respectively, where 'high' and 'low' is defined relative to a 'cut-off' energy scale Λ. Thus the ϕ_H are field oscillations of energy greater than Λ, and ϕ_L those with energy below it. The aim is now to express the path integral in these terms. Given the field variable separation, the path integral may now be written

$$Z = \int D\phi \exp\ - i \int \mathcal{L}\left(\phi, \partial_x \phi\right) d^4 x = \int D\phi_L D\phi_H \exp - i \int \mathcal{L}\left(\phi_L, \partial_x \phi_L, \phi_H, \partial_x \phi_H\right) d^4 x \tag{4}$$

We may write this in turn as

$$Z = \int D\phi_L \exp\ - i \int \mathcal{L}_{eff}\left(\phi_L, \partial_x \phi_L\right) d^4 x \tag{5}$$

Here $L_{eff}(\phi_L,\ \partial_x\ \phi_L)d^4x$ is known as the 'effective action' and is an expression involving low-energy modes only. It consists of the full Lagrangian with the higher-energy modes 'integrated out'. Now in essence, what this 'integrating out' process does is express the average of the effects of the high-energy modes on the low-energy sector of the theory. (To help make this intuitive, recall that according to the mean value theorem the average value of a function over an interval is equal to the integral of the function over that interval, divided by the interval length.) This expression thus abstracts away the differences between the individual high-energy contributions and incorporates only their net effect, as expressed in the low-energy modes. It is thus often referred to as a 'coarse-graining' of the full action $\int D\phi L(\phi)$. The question now is what the structure of this new effective action is.

The key result, which is general in scope, is that the integrating-out process generates *an infinite series of terms in the low-energy fields* $\phi_L(x)$ *in which every term consistent with the symmetry of the underlying theory is eventually* *included.*[11] Each new term comes accompanied with an undetermined constant g_i. The presence of these infinitely many undetermined parameters might raise the worry that the resultant 'theory' is predictively useless: after all, if we need to make infinitely many measurements just to determine the Lagrangian then we will never get around to using the Lagrangian to predict anything. But this worry turns out to be unfounded. For it turns out that the new couplings may, after a bit of work, be expressed in the form $g_i(E/\Lambda)^n$: that is, as functions of the energy divided by the value of the cut-off, raised to successively larger powers (see Bain 2013). A little more precisely, the new Lagrangian may be expanded in the form

$$L_\Lambda = L_0\left(\phi_{L,}\partial_x\phi_{L,}, g^*\right) + \sum_{n=1}^{\infty} g_n \left(\frac{E}{\Lambda}\right)^n O_n\left(\phi_L, \partial_x\phi_L\right) \qquad (6)$$

where L_0 is the original Lagrangian, g^* a modified value of the original coupling parameter, and the sum is over infinitely many terms in the $\phi_L(x)$ and its derivatives.[12] These terms, suppressed by factors E/Λ, are said to be *non-renormalizable.*

The Lagrangian deduced via this process therefore possesses an infinite series of terms, each with its own undetermined coupling. However, it is not unworkable from an empirical point of view, for it may immediately be seen that at energies that are very low compared to Λ—or equivalently, when we are looking at spatial scales much larger than that corresponding to the cut-off—only a very few of the non-renormalizable terms will make an appreciable contribution to the amplitude.[13] Since all prediction and measurement is done to a finite degree of empirical accuracy, this means that *only finitely many will ever need to be incorporated into calculations.* Thus L_{eff} represents a perfectly workable theory from an empirical point of view. As one field theorist, Ben Gripaios, puts it:

> If each of these operators has an arbitrary coefficient, then we need to do infinitely many measurements before we can start to make predictions. This is not a theory! We find a way out of the impasse à la George Orwell, by declaring that 'all operators are equal, but some are more equal than others'. How? Since we are interested in the physics at large-distance scales, it may be that some operators are more important at large-distances than others. This is indeed the case .... (Gripaios 2015a, p. 4)

It should be noted, however, that this effective Lagrangian is only appropriate to suitably low-energy processes. First and most obviously, since it simply has no

[11] Peskin and Schröder (1995, p. 399).

[12] The mass is shifted as well; but in QFT the mass is regarded as another coupling parameter so the general point holds.

[13] Greater energy is required to resolve small distances, meaning that energetic and spatial scales may be thought of as inverse to one another.

variables for high-energy modes it cannot be applied to compute the results of a high-energy scattering process. For another, however, as the energy approaches the cut-off Λ the influence of the previously suppressed terms grows, in the end almost certainly resulting in divergence and a violation of unitarity in addition to the loss of any predictive power.[14] Thus just from looking at equation (6) above, we can be confident that it is not valid beyond Λ, and as such that a new theory must take over.

While ϕ^4 is a simple example of a QFT, two important and general corollaries may be drawn from it.

(1) We see the beginning of an explanation of sorts of the success of the renormalization procedure. For at low energies, one effect of high-energy processes here is simply to move the value of the couplings that featured in the initial Lagrangian.

(2) We see that at low energies the effect of high-energy processes can be 'mocked up' by an infinite series of terms involving interactions of the low-energy fields. But only finitely many of these need to be considered if we are committed, as we always in fact are, to working to a finite experimental resolution.

While those are significant in themselves, the full significance of the latter point is further brought out when we consider theories with more than one field interacting. Consider for example a (toy) Lagrangian featuring two scalar fields $\Phi(x)$ and $\phi(x)$, the first heavy (mass M) and the second light (mass m), interacting as

$$\mathcal{L}_\Lambda = \frac{1}{2}\left(\partial_\mu \phi\right)^2 + \frac{1}{2}\left(\partial_\mu \Phi\right)^2 - \frac{1}{2}m^2\phi^2 - \frac{1}{2}M^2\Phi^2 - \frac{1}{2}g\phi^2\Phi. \tag{7}$$

Suppose we are interested in processes involving energies $E << M$. At these energies particle quanta of the field Φ cannot even be produced, so we are not going to see them in our accelerator. (Suppose, for example, here Φ represents the Higgs boson and we are running our accelerator prior to 2012.) The energy scale M thus represents not just some scale we happen to be interested in, but rather a real 'joint in nature' where new ontologies can be produced and new effects manifest. Nevertheless, below this energy we can ask what effect this heavy field has on what we do see. To determine that, we do as we did before and 'integrate out' the heavy field from the theory as well as all field modes at or below the energy scale M to generate an effective Lagrangian L^M_{eff}. And to first order in perturbation theory, what we find is

[14] I can't say this in good conscience without noting that in some exceptional cases the series may 'saturate' and remain well defined in the limit (see Weinberg 1995, p. 523). This is very much the exception and not the rule, however.

$$\mathcal{L}_{eff}^{M} = \frac{1}{2}\left(\partial_{\mu}\phi\right)^{2} - \frac{1}{2}m^{2}\phi^{2} + c_{0}\left(\frac{g}{M}\right)^{2}\phi^{4} + \cdots \tag{8}$$

where the ' … ' refers to operators suppressed in powers of E/M (Kaplan 1995, section 5.1). What we find, then, is the light scalar field interacting as in ϕ^4 theory plus a string of nonrenormalizable terms that will be negligible at low energies. Thus a theory that fundamentally involves two fields interacting as (7) can *look* at low energies like a theory with one field (self)interacting via (8). Again, the theory is useless above the cut-off scale M, and indeed will generally violate unitarity as it approaches it. But at low energies it is perfectly predictive provided we are interested—as we always in fact are—in finite experimental accuracy. Unlike the previous example, however, what this gives is a glimpse of how it is that the world can be fundamentally composed of a certain ontology evolving in accordance with a law of a certain structure and yet *look*, if we don't probe too carefully, as if it is composed of a different (in this case, smaller) ontology interacting via a law of a different structure.[15] As such, we get an explanation of why the physics can look significantly different from a structural point of view either side of a 'joint in nature', such as where a new particle or process comes into play.

Since the principal structural feature of laws that physicists are interested in is their symmetry structure, we should say something explicit about how the process of 'integrating out' affects symmetry.[16] The answer here is well known, and it is that *the process of integrating out preserves symmetries*. That is, the Lagrangian that is derived by this process will have all the same symmetries as the original. As van Kolck et al put it:

> The operators O_i [in (6) above] can in general be quite complicated. We can see, however, that … for an appropriate decomposition [into high and low energy fields], *they must possess all of the symmetries and transformation properties of the underlying high-energy theory*. Even if a particular symmetry is broken, it will manifest itself in the same way in the effective Lagrangian. (Kolck, Abu-Raddad, and Cardamone 2002, p. 5)[17]

[15] More complicated EFTs will permit the expression of the theory in fields taken to be bound states of the underlying ontology—for example, in chiral perturbation theory.

[16] For a review of how '[s]ymmetry considerations dominate modern fundamental physics', see Brading, Castellani, and Teh (2021).

[17] Similarly, Ecker puts the matter thus: 'To model the effective field theory at low energies, we rely especially on the symmetries of the "fundamental" underlying theory, in addition to the usual axioms of quantum field theory embodied in an effective Lagrangian. This Lagrangian must contain *all* terms allowed by the symmetries of the fundamental theory for the given set of fields (Weinberg, 1979). This completeness guarantees that the effective theory is indeed the low–energy limit of the fundamental theory' (Ecker 1995, p. 2).

Nevertheless, if we neglect the effects of the nonrenormalizable operators the low-energy theory can *look* like it has different symmetries from the original (see section 4.1 Brading, Castellani, and Teh 2021). Such symmetries arising from the neglect of small terms Weinberg termed 'accidental'. And just as embellishing an already symmetric geometric figure will tend to reduce and not extend its symmetry, the low-energy theory will generally appear to have *more* symmetry than the underlying theory. An example described by Porter Williams is of an EFT which respects the symmetries of special relativity even though the theory from which it is derived is on a discrete spacetime which strongly violates those symmetries (Williams 2019). A further example, important in the search for grand unified theories, pertains to baryon number conservation. This conservation corresponds to a symmetry of the familiar (renormalizable) Standard Model Lagrangian, and has the consequence that the proton cannot decay. It may however be shown that this symmetry is violated by certain nonrenormalizable terms of 'dimension six' and above.[18] Their high dimensionality means they must be suppressed by a factor $(E/\Lambda)^n$, with $n \geq 2$, and as such will make only a very small contribution at the energies at which we can currently test the Standard Model. Nevertheless, generic grand unified theories generically imply proton decay (such as $p \to K + \nu$), since the symmetry that results in their conservation is now considered to be merely accidental (see, e.g., Gripaios 2015b, p. 13).

What we see, then, is that QFT naturally invites the concept of hierarchies of nature in terms of a 'tower of effective field theories'. These are theories that are obtained through the process of 'integrating out' high-energy modes, and which therefore '(i) break down when pushed to scales beyond their limited domain of applicability and (ii) incorporate this inevitable breakdown into their mathematical framework' (Williams 2021). It is this outward manifestation of breakdown that justifies regarding EFTs as novel entrants into the conceptual landscape of physics—for this is not a feature exhibited by previous incarnations of non-fundamental theories. As Zinn-Justin puts it,

> the main difference between [effective] quantum field theory and non-relativistic quantum mechanics or Newtonian mechanics [is that in the latter] the mathematics doesn't tell you that it is just an approximation. Mathematically it is a fine theory. You know just from empirical evidence that it is an approximation. (Zinn-Justin 2009)

While this aspect of EFTs arguably represents something new in physics, it also gives us insight into the old problem of why it is that nature can look

[18] Dimension 5 operators are also implicated but it is dimension 6 that are now thought to have the best chance of being realized.

'radically different' at different energetic or spatial scales, and why physics is possible prior to our possession of a fundamental theory (Kolck, Abu-Raddad, and Cardamone 2002, p. 1). The simple answer is that 'the mathematical framework which we use to describe nature—quantum field theory—itself shares this basic feature of Nature: it automatically limits the role which smaller distance scales can play in the description of larger objects' (Burgess 2007, p. 330). But if this really is the explanation of the 'levels structure of theories' that physicists accept, then the theories, and the laws, that they regard as non-fundamental must be regarded as 'effective' laws generated from the more fundamental via the Wilsonian procedure. And indeed they are. For example, even the Standard Model of particle physics—our most fundamental theory to date—is regarded as an effective theory, and as such there are ongoing investigations looking for evidence of proton decay even though the corresponding terms are extremely small. It is not any empirical anomaly that leads us to invest in looking for such effects, but only our conviction that the Standard Model must be an effective theory.[19]

The fact that understanding levels in effective field theory terms gives us an explanation of why nature comes sequestered into 'levels' offers us an abductive justification for conceptualizing levels in this way. And the fact that physicists do conceptualize non-fundamental laws in this framework as effective versions of more fundamental laws suggests that we *must* understand at least some levels in this way if our metaphysics is to be extensionally adequate. But what I want to argue now is that if we do understand 'levels of laws' in these terms then these levels cannot be thought of as connected by relations of grounding. Grounding's status as the generally applicable 'level connector', and thus one applicable to the order of nature, is for that reason thrown into doubt.

4 Effective field theories and grounding

To make this claim, it will of course be necessary to say something about how grounding is understood. As anyone familiar with the literature will be aware, what partly accounts for the fact that there has been an 'explosion' of literature on grounding is the fact that almost every assumption about it has been called into question by someone. However, there are some relatively fixed points in the debate, each of which has been described as a part of the 'orthodoxy' on grounding. These include the following principles.

[19] The JUNO, Hyper-Kamiokande, and DUNE detectors are currently all searching for the signatures of proton decay events.

(1) Logic.[20] The logic of ground is a non-monotonic, strict partial order, always directed from what is less to what is more fundamental. Thus grounding is asymmetric, irreflexive, and transitive.

(2) Objectivity.[21] Grounding links up worldly entities, and what objectivity implies is that those links are themselves parts of the world existing independently of our thinking. As such, the obtaining of grounding relations is an objective fact. As Maurin puts it, 'according to the "orthodoxy" grounding is a hierarchical dependence-relation that holds between worldly facts or states of affairs. More precisely, it is an objective and mind-independently obtaining hyperintensional and non-monotonic strict partial ordering relation.'[22]

(3) Entailment.[23] Grounding is a relation of *determination*, such that the existence of a ground entails the existence of the grounded as a matter of metaphysical necessity.[24] Grounds are therefore *metaphysically sufficient conditions* of whatever is grounded in them, so that establishing that the grounds of some phenomenon are instantiated is enough to infer the instantiation of the grounded phenomenon as well.

More principles could be added to this list, but this will be enough for us to be getting along with.[25] My claim will be that the relation between successive EFTs cannot be regarded as a relation of grounding, insofar as grounding is governed by these principles. To be clear, some of the orthodox assumptions about grounding might find a happy home in the 'levels structure of theories' as conceived of within the EFT framework. In particular, the Logic requirement would seem to be satisfied.[26] The fact that successive EFTs may be defined via the same procedures checks off the transitivity requirement, and the fact that the process of 'integrating out' is 'lossy' means that the order so defined is asymmetric.[27] Rather than the Logic requirement, then, the problem for the orthodox understanding of grounding arises from a conflict between Objectivity and Entailment. And while the point I will make here is an old one—old, in any case, within the philosophy of science—it has to my knowledge yet to be raised in the context of the literature on grounding.

At the heart of the argument is the fact that EFTs—and hence, we are assuming, non-fundamental laws of nature—are by their very nature 'intrinsically

[20] Both Maurin (2019, 1574) and Rabin (2018, p. 38) describe these as 'grounding orthodoxy'.

[21] As Bliss and Trogdon (2014, section 2.1) put it, 'Grounding theorists routinely claim that grounding is fully objective.'

[22] Maurin (2019, p. 1574).

[23] Skiles (2015) notes that this is 'orthodoxy', although he himself contests it; Bliss and Trogdon (2014, section 5) call it the 'default' view.

[24] See, e.g., Rosen (2010, p. 118), who calls this 'the entailment principle'.

[25] See, e.g., Maurin (2019) for a discussion of the orthodoxy concerning grounding's relation to explanation.

[26] Or at the very least it does not raise new problems in addition to those raised below.

[27] I note also that Butterfield (2011) has argued that this process constitutes a Nagelian bridge principle, relating the languages of the high and low energy theories. In 'integrating out', we are in a sense translating the high-energy contributions into the language of low-energy fields.

approximate' entities (Castellani 2002, p. 260). In particular, even in the energy range in which they are applicable they are approximations to what is derived from more fundamental theories. This may be argued for in at least two ways.

(1) The theories that physicists use and regard as non-fundamental are predictive, empirical theories. We regard QED, for example, as a highly predictively accurate theory, and also as an EFT. But we know that theories are only predictive if they have *finitely many* undetermined constants. What is derived by the process of 'integrating out', however, is a string of terms that is *infinitely long*. As such, if we want our theory of laws to be extensionally adequate—to have what we regard as laws of physics on each side of the relation—then the laws we take to define non-fundamental levels are necessarily approximations to what is derived from the more fundamental theory.

(2) We take it that theories on different levels often have different symmetries. The whole motivation for regarding the world as structured into levels is that it appears 'radically different' on different scales, and one—and from a physics point of view, the prime—respect in which laws can differ is in terms of their symmetry structure. But we know that what is derived from a more fundamental theory must have the same symmetries as the original theory. Differences in symmetry can only arise by chopping off the Lagrangian at some point—resulting in at best an approximation of what the underlying theory entails for that scale.

It follows that the laws that physicists regard as non-fundamental are *approximations* to what may be derived from more fundamental laws (and this even in the domain in which they apply): they are approximations to the infinitely long string of terms that is derived, all but a few terms being set to zero since they will be negligible in the domain where the theory is applied. This, however, causes a familiar problem: and this is that what we take to be the non-fundamental laws are not *entailed* by those more fundamental laws. Indeed, they are generically *incompatible* with what those laws entail. This is a point familiar from some classic philosophy of science, most saliently in Feyerabend's critiques of Nagelian reduction and Hempel's deductive-nomological theory of explanation (see Feyerabend 1962, pp. 46– 7); it is also core to Duhem's criticisms of inductivism as an adequate model of Newton's method (see Duhem 1991, p. 193). In either case, the basic point is that, in addition to Newtonian mechanics providing a more comprehensive description than that provided by his predecessors (taken to be Galileo and Kepler respectively), it *corrects* what each has to say about the systems each describes (in these cases, bodies falling at the surface of the earth and planets circumnavigating the sun). But given that Newton corrects these prior theories, it cannot be that it entails them: rather, it contradicts them. Exactly the same is the case here.

Whatever the relationship between the laws on two levels, then, it cannot satisfy Entailment

There is however a response that can be made here—one which defenders of Nagelian reduction (including Nagel himself) were quick to point out in response to criticisms of Feyerabend and others. This is that they never actually intended strict deducibility as a requirement of successful reduction.[28] Rather, 'approximative reduction'—that is, derivation of an approximation to the theory that was to be reduced—is all that can and should be asked for. Nagel puts the idea as follows.

> It is undoubtedly the case that the laws derivable from Newtonian theory do not coincide exactly with some of the previously entertained hypotheses about the motions of bodies, though in other cases there may be such coincidence ... Nevertheless, the initial hypotheses may be reasonably close approximations to the consequences entailed by the comprehensive theory, as is indeed the case with Galileo's law as well as with Kepler's third Law ... But if this is so, it is correct to say that in homogeneous reductions the reduced laws are either derivable from the explanatory premises, or are good approximations to the laws derivable from the latter. (Nagel 1970, p. 120)

It is clear that Nagel regards the fact that we can derive an approximation to what is strictly derivable as sufficient to save the core of his account: we have a close enough 'analogue' to the original to say that the spirit of the original proposal is preserved. Modern apologists for Nagel's account have followed him here.[29] Such a move is clearly relevant for our purposes, since it suggests we can, without much damage, make a mild alteration to the 'grounding orthodoxy' by relaxing the requirement of Entailment to something like 'Approximate Entailment'—a principle that demands only that the more fundamental laws entail an approximation to what physicists regard as the non-fundamental laws. This, however, was a move which Feyerabend himself anticipated. As he put it:

> The objection which has just been developed—so it is frequently pointed out—cannot be said to endanger the correct theory of explanation[30] since everybody would admit that explanation may be by approximation only. This is a curious remark indeed! ... [T]he remark that we explain 'by approximation' is much too vague and general to be regarded as the statement of an alternative theory. As a matter of fact, it will turn out that the idea of approximation cannot any more be incorporated into a formal theory, since it contains elements which are essentially subjective (Feyerabend 1962, p. 48).

[28] Similar points apply to Hempel's theory of explanation.

[29] Dizadji-Bahmani et al. 2010, Butterfield 2011.

[30] Here he has in mind Nagel's account as well as Hempel's deductive-nomological model—models which he regarded to 'not differ in any essential respect' (Feyerabend 1966, p. 247).

Feyerabend's objection here, then, is that if we weaken 'derivability' to mean 'approximate derivability', we (i) produce a theory which is so vague that it cannot be stated, and (ii) sacrifice its objectivity. While I regard (i) as untrue—or at least as unfair, given that vagueness permeates most of our theoretical notions—the objection in (ii) remains absolutely correct.[31]As such, while the modification of Entailment required to deal with the fact that what we regard as non-fundamental laws are 'intrinsically approximate' might save the spirit of that principle, it does so at the cost of sacrificing Objectivity—another central tenet of the orthodoxy on grounding.

To see this, let's start with the claim that 'approximation' is so 'vague and general' that to include it in a theory of explanation is essentially to abandon one's ambition in providing a theory. Here defenders of Nagelian reduction have argued that while there may be no general philosophical theory that one can offer as to when two equations, or two theories—indeed, two anythings—are sufficiently similar to be regarded as 'approximations' of one another, we can still give content to the claim. It is just that the content is invariably going to be contextual. And in any empirical context, to say a successor theory approximates a precursor theory can be expected to at least involve the claim that the two theories are *approximately empirically equivalent* in the domain in which the old theory proved successful. Of course, what that means, and thus whether it is true, is going to depend on facts about the context of investigation. Once that context is specified, however, the truth value may be determined unambiguously.

So the problem afoot here is not 'vagueness'. Rather, the real problem is that weakening Entailment to something like Approximate Entailment steers us into the second horn of Feyerabend's dilemma, in that it implies a conflict with Objectivity. Feyerabend's own stated reasons for regarding approximation as inevitably involving 'elements which are essentially subjective' turn on considerations of theoretical incommensurability which are difficult to summarize, and probably in any case a bit dated. However, we can turn to a classic discussion by Duhem to see more clearly why the point stands. In this discussion, he first argues on broadly empiricist grounds that 'every physical law is an approximate law' (Duhem 1991, p. 171). From there, he writes:

> Such [an approximate] law [is] always provisional ... It is provisional because it represents the facts to which it applies with an approximation that physicists today judge to be sufficient but will some day cease to judge to be satisfactory. Such a law is always relative; not because it is true for one physicist and false for

[31] Thus while modern apologists for Nagel's account argue against the idea that the notion of 'approximation' between theories is devoid of content, holding that what that content is is nevertheless highly 'contextual' (see, e.g., Dizadji-Bahmani, Frigg and Hartmann 2010). I agree; I hold further that those contexts involve 'subjective elements'.

> another, but because the approximation it involves suffices for the use the first physicist wishes to make of it and does not suffice for the use the second wishes to make of it …
>
> The estimation of its value varies from one physicist to the next, depending on the means of observation at their disposal and the accuracy demanded by their investigations. (Duhem 1991, p. 171)

What Duhem is pushing here is that whenever a law is regarded as in some way approximate, *whether it can be regarded as a law at all* is not only a 'contextual' matter but moreover one that *depends on the relevant interests*. Of course, we need not buy into Duhem's explicitly empiricist motivations for believing that all laws are necessarily approximate to take this conclusion seriously: all that is needed to generate the problem is that this is true of the non-fundamental laws corresponding to effective field theories. But such theories, as has been emphasized above, *are* inevitably and intrinsically approximate. And whether an EFT containing n terms and exhibiting symmetry S constitutes a good approximation to what can be strictly derived from the more fundamental theory is going to depend upon what we are interested in studying and the degree of accuracy with which we are interested in studying it. As such, Duhem's point applies here. It follows from this that the relata of the relation connecting successive EFTs are interest-dependent entities. Since I take it as uncontroversial that a relation can obtain in an interest-independent sense only if all of its relata obtain in that way, *the link between the laws given by effective theories cannot be identified with grounding*.[32]

It may help to flesh this out with an example already alluded to. As noted above, for the purposes of most particle physicists the Lagrangian of the Standard Model, applied at some energy scale E, where E is below Λ and hence in the range in which it is defined, is just the plain old renormalizable Lagrangian of the Standard Model—the one that can be found displayed on mugs and T-shirts in the CERN gift shop. But *for those interested* in studying proton decay, and experimentally well positioned to do so, this is not the Lagrangian that is appropriate: dimension 6 corrections to the Standard Model, of the form $O(E/\Lambda)^2$, must be included if such a phenomenon (should it exist) is to be accounted for. It is important to note that the issue here is not simply that 'different things happen at different levels' with levels parceled out at different energies or spatial scales—perhaps analogously to how we see different thermal phenomena, such as freezing or boiling, happening at different temperature scales. For we can hold the energy range—the 'scale'—fixed and *still* ask whether the higher-order terms can be neglected or not; and the answer will depend on our

[32] As Bliss and Trogdon (2014, section 2.1) note, given the variety of meanings associated with the word 'objective', there are a number of different ways in which grounding could fail to be objective. This way seems to correspond to the relata of the grounding relation being 'essentially connected to subjects', and thus to what they call the 'more "metaphysical" approach' to the failure of objectivity.

interests. To put the same point differently, we should not think it is at a scale 'beyond' the domain at which it is first appropriate to ascribe protons and the more familiar hadronic decays processes at which proton decay kicks in. For the smallness of the higher-order terms corresponding to the decay of the proton simply translates (via cross-sections and decay rates) into the extreme *rareness* of the proton decay events, relative to other such decays. Thus if we will detect proton decay, we will do so in broadly the same sorts of detectors we use to see many other hadron decay events (a well-shielded chamber of fluid surrounded by array of detectors); we do not need a detector that probes to 'deeper' scales so much as just a particularly *capacious* such detector, kept track of for sufficiently many years.[33] Thus it seems to me that proton decay happens at just the same 'scale' as more familiar hadron decay processes.[34] But it depends on one's interests whether the corresponding terms, evaluated at that scale, may or may not be set to zero. As such, what law is 'approximately entailed' at a given energy by the more fundamental successor to the Standard Model is a function of the interests of the physicist, and Duhem's point stands.

To summarize the argument of this section. A core principle of the grounding orthodoxy is that the grounds entail what they ground. But the laws of nature that we take to be described by effective theories are not entailed by the more fundamental theory: rather, they contradict what is entailed. And if we modify Entailment to mean something like 'Approximate Entailment', as defenders of the Nagelian model think we should, then this modification is in contradiction with the idea that grounding relations satisfy Objectivity. For the very relata of the intertheory relation are not objectively determined, even at a particular 'scale': on the contrary, what the laws at a given scale are is a function of the interests of the theorist. Again, since I take it as uncontroversial that a relation can only obtain in an interest-independent sense if all of its relata obtain in that way, it cannot be that the link between the laws described by effective theories is identifiable with grounding. And since I take a level to be 'a domain with its own set of entities, structures, and laws', levels themselves are not connected by relations of grounding either.

5 Responses

Today's physicists generally understand levels in terms of a tower of effective field theories, linked via Wilsonian methods; today's metaphysicians generally

[33] As the Hyper-Kamiokande project webpage puts the matter: 'With the giant detector, data that would take 100 years to obtain with Super-Kamiokande can be obtained in about ten years with Hyper-Kamiokande. This makes it possible to measure rare phenomena of elementary particles and slight symmetry breaking that were previously invisible.' (Hyper-Kamiokande, 2024).

[34] If it is useful to make an analogy here, consider a very rare genetic condition that results in a very rare genetic transcription process. We wouldn't say that the rare transcription process took place on a different 'scale' from other such processes simply because it was rare.

understand levels in terms of relations of grounding. I have argued that the relations of grounding do not correspond to the relations between levels in the physicists' sense. Given how pervasive is the belief in metaphysics that grounding is involved whenever one talks about a world structured into 'levels', I expect some pushback against the argument just given. While there are of course a number of objections that could be made to that argument, here I mention just three.

A first objection is that the argument just given applies only to laws of nature. But there are other relata of the hierarchy of nature—most saliently, that of more and less fundamental objects and that of more and less fundamental properties—that are not directly touched by the argument. A full response to this objection would have to say more that is explicit about how more and less fundamental ontology is to be understood in the EFT framework. Suffice to say here however that I am with Nagel when he writes that since the objects and properties described by our theories are just that—described by our theories—we need to understand the relations between those through the prism those theories provide (Nagel 1961, p. 270). Since laws are at the core of scientific theories and what we do with them, I do not think my argument will simply go away even if we do focus on a hierarchy different from the hierarchy of laws.

A second objection is that one could say that it is the entire infinite string of operators derived through Wilsonian methods that corresponds to a non-fundamental level, and what this is is not a function of anyone's interests. Thus, relations of grounding do obtain between levels after all. However, as noted above in Section 4, this string does not in fact correspond to anything that physicists call a 'law' (as the earlier quote from Gripaios puts it, 'this is not a theory!'). Moreover, this 'law' has the same symmetries as the underlying theory, and thus is not the law of any level with different symmetry structure than the fundamental theory. A level, by contrast, I am taking to be 'a domain with its own set of entities, structures, and laws'. For both these reasons, this infinitely long string does not correspond to the law of a non-fundamental 'level', and so the relation between it and the fundamental theory is not the relation between levels that we are looking for.

A third and perhaps most important objection is that the argument overgeneralizes. For it is not as though it is only the laws as they appear in the framework of effective theories that are both non-fundamental and approximate (see, e.g., Callender 2001). Indeed, one can easily find the terms 'non-fundamental law' and 'approximate law' used interchangeably in the literature. As such, there is at best nothing new in this objection that the hierarchy of laws is not objective. And at worst, since approximation and idealization are utterly ubiquitous in scientific practice, practically nothing in science is going to come out as objective by my measure; who cares, then, that 'levels' and 'grounding' do not do so either.

In one sense I agree with this: I was after all explicit that I am going over old ground here. But for all that, I think that there is something both new and newly consequential for metaphysics here. After all, it is the EFT framework that is taken to *explain* why it is that nature admits of something like a levels structure. Prior to EFTs, one was freer to understand that approximate character of non-fundamental laws in terms of something about human fallibility: something that, while revealing about humans, did not necessarily undermine the objective reality of non-fundamental laws. Now the situation is different. We find ourselves in a situation in which the 'hierarchy of levels' that is now not just *reported* but *explained* by physics turns out not to be fully specifiable in an interest-independent way, even if it also makes room for objective 'joints in nature' such as at particle masses. Thus it is not *wrong* for those working in 'stratified metaphysics' to take as their starting point ideas about a 'hierarchy in nature', which as I flagged at the outset is standard practice. The above should, however, give metaphysicians permission to think about levels in less committedly realist terms. I suspect that talking in terms of 'grounding' is only going to hinder this endeavour, given the pervasiveness of the assumption that grounding obtains independently of us. For what seem to be fundamental reasons, this now does not seem to be fully apt.

References

Bain, J. (2013). "Effective field theories". In: *The Oxford Handbook of Philosophy of Physics*. Ed. by R. Batterman. Oxford University Press, p. 224.

Bhogal, H. (2017). "Minimal anti-humeanism". *Australasian Journal of Philosophy* 95.3, pp. 447–460.

Bliss, R., and Trogdon, K. (2014). "Metaphysical grounding". In: *Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University.

Brading, K., Castellani, E., and Teh, N. (2021). "Symmetry and symmetry breaking". In: *The Stanford Encyclopedia of Philosophy*. Ed. by E. N. Zalta. Fall 2021. Metaphysics Research Lab, Stanford University.

Burgess, C. P. (2007). "An introduction to effective field theory". *Annu. Rev. Nucl. Part. Sci.* 57, pp. 329–362.

Butterfield, J. (2011). "Emergence, reduction and supervenience: a varied landscape". *Foundations of Physics* 41, pp. 920–959.

Callender, C. (2001). "Taking thermodynamics too seriously". *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32.4, pp. 539–553.

Cao, T. Y., and Schweber, S. S. (1993). "The conceptual foundations and the philosophical aspects of renormalization theory". *Synthese* 97, pp. 33–108.

Castellani, E. (2002). "Reductionism, emergence, and effective field theories". *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 33.2, pp. 251–267.

Craver, C. (2015). "Levels". *Open MIND* 8(T), pp. 9–10. doi: 10.15502/9783958570498.

Dizadji-Bahmani, F., Frigg, R., and Hartmann, S. (2010). "Who's afraid of Nagelian reduction?" *Erkenntnis* 73, pp. 393–412.

Duhem, P. M. M. (1991). *The Aim and Structure of Physical Theory*. Vol. 13. Princeton University Press.

Ecker, G. (1995). "Chiral perturbation theory". *Progress in Particle and Nuclear Physics* 35, pp. 1–80.

Feyerabend, P. K. (1962). "Explanation, reduction, and empiricism". In: *Scientific Explanation, Space, and Time, Minnesota Studies in the Philosophy of Science, Volume III*. Ed. by H. Feigl and G. Maxwell. University of Minnesota Press, pp. 103–106.

Feyerabend, P. K. (1966). “The structure of science”. *British Journal for the Philosophy of Science* 17.3, pp. 237–249.

Gripaios, B. (2015a). “Lectures on effective field theory”. *arXiv preprint arXiv:1506.05039*.

Gripaios, B. (2015b). “Lectures on effective field theory”. *arXiv preprint arXiv:1506.05039*.

Hartmann, S. (2001). “Effective field theories, reductionism and scientific explanation”. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32.2, pp. 267–304.

Hyper-Kamiokande. (2024). “About: Physics,” https://www-sk.icrr.u-tokyo.ac.jp/en/hk/about/research/. University of Tokyo.

Kaplan, D. B. (1995). “Effective field theories”. *arXiv preprint nucl-th/9506035*.

Kolck, U. van, Abu-Raddad, L. J., and Cardamone, D. M. (2002). *Introduction to effective field theories in QCD*. *arXiv*: nucl-th/0205058[nucl-th].

Kovacs, D. M. (2017). “Grounding and the argument from explanatoriness”. *Philosophical Studies* 174.12, pp. 2927–2952. doi: 10.1007/s11098-016-0818-9.

Ladyman, J., and Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized*. Oxford University Press.

Maurin, A. (2019). “Grounding and metaphysical explanation: it’s complicated”. *Philosophical Studies* 176.6, pp. 1573–1594.

McKenzie, K. (2022). *Fundamentality and Grounding*. Cambridge University Press.

Nagel, E. (1970). “Issues in the logic of reductive explanations”. In: *Mind Science and History*. Ed. by H. E. Kiefer and M. K. Munitz. State University of New York Press, pp. 117–137.

Nagel, E. (1961). *The Structure of Science: Problems in the Logic of Scientific Explanation*. Vol. 1. Harcourt, Brace & World New York.

Potochnik, A. (2017). *Idealization and the Aims of Science*. University of Chicago Press.

Peskin, M., and Schroeder, D. (1995). *An Introduction to Quantum Field Theory*. CRC Press.

Rabin, G. (2018). “Grounding Orthodoxy and the Layered Conception”. In: *Reality and its Structure: Essays in Fundamentality*. Ed. by R. Bliss and G. Priest. Oxford University Press, pp. 37–49.

Rivat, S., and Grinbaum, A. (2020). “Philosophical foundations of effective field theories”. *European Physical Journal A* 56, pp. 1–10.

Rosen, G. (2010). “Metaphysical Dependence: Grounding and Reduction”. In *Modality: Metaphysics, Logic and Epistemology*. Oxford University Press, pp. 109–135.

Schaffer, J. (2009). “On what grounds what”. In: *Metametaphysics: New Essays on the Foundations of Ontology*. Ed. by D. Manley, D. J. Chalmers, and R. Wasserman. Oxford University Press, pp. 347–383.

Schaffer, J. (2016). “Grounding in the image of causation”. *Philosophical Studies* 173.1, pp. 49–100. doi: 10.1007/s11098-014-0438-1.

Sider, T. (2020). “Ground grounded”. *Philosophical Studies* 177.3, pp. 747–767. doi: 10.1007/s11098-018-1204-6.

Skiles, A. (2015). “Against Grounding Necessitarianism”, *Erkenntnis* 80.4, pp. 717–751.

Van Kolck, U., Abu-Raddad, L., and Cardamone, D. (2002). “Introduction to effective field theories in QCD”. *AIP Conference Proceedings* 631.1. American Institute of Physics, pp. 191–220.

Weinberg, S. (1979) “Phenomenological Lagrangians”, *Physica A: Statistical Mechanics and its Applications* 96.1, pp. 327–340.

Weinberg, S. (1995). *The Quantum Theory of Fields*. Vol. 2. Cambridge University Press.

Weinberg, S. (2021). “On the development of effective field theory”. *European Physical Journal H* 46, pp. 1–6.

Williams, P. (2019). “Scientific realism made effective”. *British Journal for the Philosophy of Science* 70.1, pp. 209–237.

Williams, P. (2021). “Renormalization group methods”. *The Routledge Companion to Philosophy of Physics*. Ed. by E. Knox and A. Wilson. Routledge, pp. 296–310.

Zinn-Justin, J. (2009). *Asymptotic Safety: A Review*. url: https://pirsa.org/09110041/. Perimeter Institute.

PART V

LEVELS OF EXPLANATION IN MATHEMATICS AND METAPHYSICS

13

Explanation in Descriptive Set Theory

Carolin Antos and Mark Colyvan

1 Introduction

We are interested in explanations in mathematics. These are sometimes called *intra-mathematical explanations* and involve one mathematical result being explained in terms of further mathematics. For example, some proofs are explanatory: they do more than merely show *that* a given theorem is true; they demonstrate *why* the theorem is true. It is an interesting, open question whether explanation in mathematics is always connected with a proof of a theorem. While there is good reason to suspect that proof may not be the only locus of explanation in mathematics,[1] it is, at least, *one* such locus. For present purposes, we set aside the issue of non-proof-based explanation and concentrate on explanations arising from proofs.

Explanation in mathematics is important for a number of reasons. For a start, such explanation is clearly not causal so is not accommodated by causal accounts of explanation, such as those advanced by Lewis (1986). At least as traditionally construed, mathematical facts are necessarily true, so counterfactual accounts of explanation run into trivialism issues when applied to mathematics.[2] In light of all this, mathematical explanation is an interesting test case for theories of explanation and presents problems for any ambitions for a single, unified theory of explanation (see Reutlinger et al., 2022). Of course, mathematical explanation is interesting in its own right. In this paper, we are less interested in the broader philosophical issues[3] and more concerned with understanding mathematical explanation in its own terms.

[1] See, for example, Colyvan (2012); Lange (2018); D'Alessandro (2020).

[2] The core idea of a counterfactual account of explanation is that *A* explains *B* just in case the following counterfactual holds: had *A* not been the case, then *B* would not be the case. But in mathematics, both *A* and *B* are necessary, so the counterfactual in question has an impossible antecedent so is trivially true (at least, according to the usual semantics for counterfactuals). There has been some work on extending such counterfactual accounts to deal with mathematical explanation by invoking counterpossibles (see Baron et al. 2020).

[3] See Mancosu (2008); Colyvan (2018) for more on this.

Carolin Antos and Mark Colyvan, *Explanation in Descriptive Set Theory* In: *Levels of Explanation.* Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Carolin Antos and Mark Colyvan 2024.
DOI: 10.1093/oso/9780192862945.003.0014

In pursuit of this goal, it is instructive to look at theorems that have different styles of proofs. In particular, it is good to look at explanatory proofs and non-explanatory proofs of a particular result. The fact that such pairs of proofs exist for some theorems helps establish that it is not the theorem itself that is explanatory or not. The explanation seems to reside in at least some proofs. Moreover, looking at such pairs of proofs allows us to identify the explanatorily relevant differences between the proofs and thus help identify what makes a proof explanatory. This, in turn, helps us get a grip on what a theory of explanation in mathematics might look like. Also of interest are pairs of proofs of a particular theorem where each has some claim to being explanatory but in different ways (or perhaps at different levels of generality).

Thus far, the philosophical literature on mathematical explanation has mostly focused on examples of proofs from elementary number theory, Euclidian geometry, and the like. Focusing on such basic mathematics is understandable. Indeed, it is usually advisable to use as simple an example or case study as is needed for the task at hand. And, of course, examples from elementary mathematics will be accessible to a broader range of readers. The problem with this, however, is that we run a risk of developing an account of explanation that is based on too limited a stock of examples and does not do justice to mathematics as a whole. We think it is important to draw examples from different areas of mathematics. We hold this view for a couple of reasons. First, if we focus too narrowly on elementary examples, we might be misled about the nature of mathematical explanation in higher mathematics. We need at least some examples from advanced mathematics. Second, there may well be different explanatory goals and even different standards of proof in different areas of mathematics. We thus must consider examples from at least some of the many different branches of mathematics. Ideally, we would draw examples from across all areas of mathematics. This is impractical in a paper such as this. Instead we see this paper as a contribution to this larger task: the diversification of examples needed for informed philosophical discussion about mathematical explanation.[4] Our main focus will be on proofs in one advanced area of modern mathematics—descriptive set theory—where there has been some very interesting debate over mathematical explanation in dichotomy theorems.

Finally, we note that we need to draw on the judgements of mathematicians about which proofs are explanatory if we are to respect mathematical practice here. It is all too easy for philosophers' judgements about which proofs are explanatory to be clouded by other philosophical commitments (e.g., in metaphysics, in epistemology, and in the philosophy of explanation). In a naturalist spirit, we see our task to be that of taking the judgements of mathematicians and trying to make philosophical sense of these.

[4] There has already been some work in this direction, e.g., Lange (2017); Colyvan et al. (2018).

2 Fermat's Little Theorem

Before we get to our main case study in descriptive set theory, it will be useful to warm up with an elementary example. This will help to get a feel for the issues in question. The example of this section is from number theory and is known as *Fermat's Little Theorem*.[5]

Theorem 1 (Fermat's Little Theorem). *If p is prime and a is a positive integer such that $p \nmid a$ (p does not divide a), then* $a^{p-1} \equiv 1 \pmod{p}$.

There are many different proofs of this theorem. Arguably many of these proofs give different insights into the theorem and forge connections with different branches of mathematics. Here we're content to provide sketches of three different proofs and make some suggestions about their relative explanatoriness.

2.1 A Number Theory Proof

Consider the set of $p-1$ integers $S = \{a, 2a, 3a, \ldots (p-1)a\}$. None of these integers is divisible by p, for if $p \mid ja$ (i.e., p divides ja) for $1 \leq j \leq (p-1)$, then, since $p \nmid a$, we'd have the impossibility: $p \mid j$. Moreover, no two of the integers in S are congruent modulo p. If they were, we'd have j and k less than $p-1$ such that $ja \equiv ka(\text{mod } p)$. But since $p \nmid a$, this means that $j \equiv k(\text{mod } p)$, but this is impossible since both j and k are less than $p-1$. This means that the least positive residues (modulo p) of the members of S are the integers $1, 2, 3, \ldots (p-1)$. So $a \cdot 2a \cdot 3a, \ldots (p-1)a \equiv 1 \cdot 2 \cdot 3 \ldots (p-1) \pmod{p}$. Thus $a^{p-1}(p-1)! \equiv (p-1)! \pmod{p}$. Since $(p-1)!$ and p are relatively prime, we can divide both sides of the last equivalence by $(p-1)!$ to give us the required result: $a^{p-1} \equiv 1(\text{mod } p)$.[6]

This proof uses only number-theoretic resources and has some claim to being explanatory. It shows that the result holds because of facts about prime numbers, divisibility, and the like. In essence, we have a number-theory result spelled out in terms of the properties of numbers. Such proofs are valued in number theory and are called 'elementary proofs' and are contrasted with some proofs that proceed via methods from complex analysis. It is not clear that elementary proofs in number theory are valued for their explanatoriness, but this is a fair assumption. After all, if we have a theorem about prime numbers, we could reasonably expect that an explanation of this would be in terms of properties of

[5] Thanks to Hannes Leitgeb for suggesting this example and for his mathematically informed intuitions about which of the proofs is more explanatory.

[6] A version of this proof can be found in Rosen (1988).

prime numbers—not rely on facts about analytic functions on the complex domain.[7] This proof fits the bill and seems to give us insights into why the theorem holds. But can we do better?

2.2 A Group Theory Proof

It is straightforward to show that $G=\{1,2,3,\dots p-1\}$, with the operation of multiplication $(\text{mod } p)$, is a group. Next we reduce a modulo p so we can assume that $1 \leq a \leq p-1$. That is, $a \in G$. Let k be the smallest positive integer such that $a^k \equiv 1(\text{mod } p)$. Then the set containing the numbers $1, a, a^2, \dots a^{k-1}$, reduced modulo p, forms a subgroup of G with order k. We then invoke Lagrange's Theorem[8] to show that k divides $p-1$ (which is the order of G). So we have $p-1=kn$, for some positive integer n. Thus $a^{p-1} \equiv a^{kn} \equiv 1^n \equiv 1(\text{mod } p)$, as required.[9]

This proof shows that Fermat's Little Theorem is an instance of a more general group-theoretic result. At least, the proof places this number-theoretic result in a broader context of group theory. Indeed, Lagrange's theorem is the key to this particular proof. It is worth noting that there are group-theoretic proofs that do not invoke the full generality of Lagrange's theorem but, instead, prove the crucial step directly by proving, in effect, a special case of Lagrange's theorem.[10] It is the generality delivered by this proof that gives it its claim to explanatoriness. While Fermat's Little Theorem is a number-theoretic result, this group theory proof, we think, is more explanatory. But this does raise an interesting question about whether it is generality that matters most or proving a result in a particular area by appealing to details of the area in question.[11] The number theory proof in the previous section had the latter virtue. We might think of this earlier proof as delivering a local or intrinsic notion of explanation, while the group theory proof offers a more unifying or global notion of explanation. Indeed, these might be thought of as distinct axes of evaluation of a proof, both relevant in their own right, but not always offering a best balance between the two. We will return to such issues in our discussion of the major case study presented in the next section.

2.3 A Combinatorial Proof

As before, assume that p is prime and a is a positive integer such that $p \nmid a$. Suppose we have a different coloured beads and we wish to make necklaces with p

[7] The preference for elementary proofs in number theory resonates with Hartry Field's (2016) argument preferring *intrinsic explanation* in science.

[8] This theorem states that the order of any subgroup of a finite group G divides the order of G.

[9] A version of this proof can be found in Weil and Rosenlicht (1979).

[10] Euler provided such a proof (1761).

[11] This raises the interesting question of whether levels of generality might correspond to levels of explanation.

beads in each. First we place p beads on a string and we note that there are a^p such possible strings. We discard the strings consisting of beads of only one colour. This leaves $a^p - a$ strings. Now we join the ends of the strings to form the necklaces. We note that some of the necklaces are cyclic permutations of others. While the cyclic permutations are distinguishable as untied strings, they are indistinguishable as necklaces. Since there are p cyclic permutations of the beads on the string, and p is prime, the number of distinguishable necklaces is $(a^p - a) / p$ and this must be an integer. From this the result follows.[12]

This is an interesting proof. It uses the least sophisticated mathematics: there's no group theory or even much by way of number theory here. Moreover, the appeal to necklaces helps with visualisation. Indeed, the proof has the reader build a mental model in the service of delivering the result in question. For these reasons, this proof is very useful pedagogically. It is explanatory in the sense that it helps newcomers to number theory get a grip on Fermat's Little Theorem. But it also has some claim to being explanatory in the sense of revealing the real reason that the result holds. After all, the construction of necklaces and discarding of duplicates is, in a sense, just some do-it-yourself group theory. Or rather, what we have here is a particular instance of the group-theoretic approach but without invoking the full generality of group theory. There is no appeal to groups or Lagrange's theorem to be seen in this proof, yet it is a particular instance of Lagrange's theorem, applied to the case at hand, that lies at the heart of this proof. So this proof might be thought to have many of the virtues of the group theory proof but without the full generality afforded by group theory. This proof thus does not (explicitly, at least) forge a connection between number theory and group theory. For this reason, it might be argued that this proof is, indeed, explanatory but perhaps not as explanatory as the more unifying group theory proof.

Nothing hangs on our tentative suggestions about the relative explanatoriness of the above three proofs. We simply note that if these proofs all have some claim to explanatoriness, arguably, it is for different reasons. Moreover, it seems that explanatoriness comes in degrees; we are not dealing with an all-or-nothing concept here.

3 Dichotomy Theorems in Descriptive Set Theory

This case study is from descriptive set theory, a sub-area (or perhaps even a neighbouring area) of set theory that is more strongly connected to standard mathematics than the more abstract, logical areas of common set theory. Here we will outline an example from recent descriptive set theory that showcases different

[12] This proof can be found in Goloumb (1956).

aspects of explanatoriness in the proofs of a class of theorems: *the dichotomy theorems*. This class encompasses a number of theorems which are based on a classical result by Cantor and then generalised to ever more abstract levels. Our focus here lies on the existence of two main proof types for these theorems, where the introduction of the second proof type signified a strong discontinuity in proving such theorems. In the following we will give the argument that both proof types present us with elements that make them explanatory, albeit in very different ways.

The main aim of giving this case study is to present an example about explanatoriness from very recent research, something that is missing in the literature on mathematical explanation. We think that such an example can shed further light on the complexities of mathematical explanation and its impact on recent research. In particular, our example will show that the search for explanatoriness is a major motivating factor for producing new and fruitful research, leading to fundamental discussions in the expert community and influencing the direction of research. Furthermore, we will show how a type of pluralism in explanatoriness can occur that is related to different sub-areas in mathematics and their respective communities.

Studying such an advanced example brings some peculiarities in presentation as well as content. We will usually not be able to give the whole proofs under consideration or even a detailed outline of them, as the mathematical background theory is too technical and would involve more setting up than we can accomplish here. Instead, we will present the main mathematical intuitions behind the results in question, leaving the details to textbooks and articles on the subject. However, we think that these limits in exposition are compensated by some unique insights with which examples from recent research provide us.

One advantage is the possibility to observe current discussions about explanatoriness and related questions by mathematicians themselves. We can see this more clearly in recent research because we have access not only to the formalised content as presented in textbooks or papers but also to informal material such as slides from talks, programmatic research statements, and discussions with the mathematicians themselves.[13]

Further, when considering recent research we are often presented with a far more complex and advanced mathematical setting, making it hard for an average investigator into explanatoriness to develop an intuition of her own that goes beyond reconstructing the reasoning of the experts. Here we have to solely rely on the intuition about explanatoriness of the mathematicians in the relevant field, thus making our approach more independent from our own views on the matter. The complexity of recent mathematical research can also highlight problems with accounts of explanatoriness that are not so clearly seen when considering

[13] For historical case studies, similar things can sometimes be found in correspondences or in cases the theorems are especially surprising (for such a case study, see Hafner and Mancosu, 2005).

examples from more elementary mathematics. One instance for this is that of the explanatoriness of parts of proofs. For example, one could ask whether for a proof to be explanatory, do all the proofs of all the lemmata, theorems, or basic facts that are used in it have to be explanatory as well? Typical examples from contemporary mathematics will make issues such as these more pressing, as they usually rely on a widespread network of existing mathematical results.

Descriptive set theory is a part of set theory that studies definable subsets of the real numbers in certain topological spaces. Although it is considered to be a subdiscipline of set theory, it is also connected to more mainstream mathematics—areas such as topology and functional analysis.

Dichotomy theorems are a class of theorems that go back to the beginnings of (descriptive) set theory. Indeed, as with so many things in set theory, the earliest version of such a dichotomy theorem arose in the works of Cantor when searching for a solution to the Continuum Hypothesis (CH), the hypothesis that there are no infinite cardinals strictly between the size of the natural and the real numbers. One partial result by Cantor (1884) implies that the CH holds for closed sets, i.e., sets that contain all of their limit points. This was the starting point for a line of theorems, the set-theoretic dichotomy theorem (they state either-or results), that generalise Cantor's result step by step by considering ever more abstract definable subsets of the real line. Together they provide detailed insight into the mathematical structure of the continuum and constitute one of the core areas of descriptive set theory.

Following the exposition of the history of set-theoretic dichotomy theorems provided in Miller (2012), we can see that the continuity in ever more general versions of set-theoretic dichotomy results did not transfer to the proof structure of these theorems. Instead, Miller (2012) points to a strong discontinuity between the proof of Cantor's theorem and early generalisations to Borel and analytic sets,[14] on the one hand, and, on the other hand, the proofs of a later generalisation by Silver (1980) and subsequent work. There is a proof type for the earlier theorems that has a mathematical construction at heart that is considered to be especially informative (we will call this *the classical proof type*). For the later theorems such a type of proof was not available for some decades. Instead, these proofs relied on very advanced techniques from other areas of mathematical logic, in particular from recursion theory and general set theory (we will call this *the advanced logic proof type*). Only very recently was B. Miller able to find a proof that relies on comparable principles as the one for the older theorems (see, for example, Miller 2011). For a partial timeline of the theorems and proofs, see the chart below:

[14] The definitions will be provided in the next sections, when the relevant theorems are considered in more detail.

Year	Dichotomy Theorems	Proof Type
1884	Cantor-Bendixon	Classical
1916	Hausdorff/Alexandroff	Classical
1917	Souslin	Classical
⋮	⋮	⋮
1980	Silver	Advanced logic
⋮	⋮	⋮
1990	Harrington-Kechris-Louveau	Advanced logic
1999	Kechris-Solecki-Todorcevic (KST)	Advanced logic
to the present	⋮	⋮
2010	KST, Silver etc.	(New) classical
to the present	⋮	⋮

As we will see, that classical proof type relies on the construction of so-called Cantor-Bendixon derivatives (see Definition 2). The new classical proof scheme uses a similar construction while at the same time forgoing the use of advanced logical techniques that were introduced for the original proof of Silver's theorem. In the literature, the classical and new classical proof schemes are therefore considered as one type of proof and the advanced logic proof as another.

In the following we will analyse these two types of proofs with respect to their explanatory value. As the main arguments for the explanatoriness of the proof types often involves several of the dichotomy theorems or the interrelations between them, we will mostly consider (parts of) the class of dichotomy theorems instead of one single theorem.[15]

3.1 The Classical Proof Schema

3.1.1 Early Dichotomy Theorems

The classical proof schema goes back to the first version of a dichotomy theorem related to Cantor's result in Cantor (1884). We will start by considering this example in greater detail, as it provides the basic construction that is used in the classical proof schema: 'We can think of the Cantor-Bendixon Theorem as a construction principle, since it gives us a method of building up the closed sets from the apparently simpler perfect sets and countable sets' (Moschovakis, 2009, 51).

[15] However, there are a few that come up more often, because of their general significance. Amongst these are the Cantor-Bendixon theorem Cantor (1884), Souslin's theorem for analytic sets (1917), and Silver's theorem (1980).

Definition 1.[16]

- *A space is* perfect *if all its points are limit points. If P is a subset of a topological space X, we call P* perfect in *X if P is closed and perfect in its relative topology.*
- *A point x in a topological space X is a condensation point if every open neighbourhood of x is uncountable.*

Theorem 2 (Cantor-Bendixon). *Let X be a Polish space (i.e., a separable completely metrisable space). Then X can be uniquely written as* $X = P \cup C$, *with P a perfect subset of X and C is countable and open.*[17]

This theorem can be proven in a quite simple manner, where we provide a construction of the partition of *X*: $P = \left\{ x \in X : x \text{ is a condensation point of } X \right\}$ and $C = X \setminus P$.[18]

However, there is also a more general construction mechanism for the perfect set. The idea is that the perfect set we are looking for is a specific set in a decreasing transfinite sequence of closed subsets of the space *X*. The definition goes as follows:

Definition 2. *For any topological space X, let*

$$X' = \{x \in X\colon x \textit{ is a limit point of } X\}.$$

We call X' the Cantor-Bendixon derivative of X. Then X' is closed, X is perfect if and only if X = X'. Repeating this definition transfinitely many times gives rise to the following construction: let X^α be the iterated Cantor-Bendixon derivatives for all ordinals α, defined as follows:

(1) $X^0 = X$,

(2) $X^{\alpha+1} = \left(X^\alpha\right)'$,

(3) $X^\lambda = \bigcap_{\alpha < \lambda} X^\alpha$, if λ is limit.

Then $(X^\alpha)_{\alpha \in ORD}$ *is a decreasing transfinite sequence of closed subsets of X.*

It can now be shown that the perfect kernel *P* of the Cantor-Bendixon Theorem is X^{α_0}, where α_0 is a countable cardinal for which $X^\alpha = X\alpha_0$ for all

[16] Unless marked otherwise, the following definitions, theorems, and proofs are taken from Kechris (1995). See there for more details and background.

[17] To better see the connection with the later theorems, consider a different version of this theorem. Suppose that *X* is a Polish space and $C \subset X$ is closed. Then exactly one of the following holds: either *C* is countable or there is a perfect subset of *C*.

[18] It remains to show the desired properties of *P* and *C* and prove the uniqueness of the partition; see Kechris (1995, 32).

$\alpha \geq \alpha_0$(that such an α_0 exists follows from a more general fact about specific descending sequences).[19]

3.1.2 The Explanatory Value of the Classical Proof Type

It is interesting to note how these two proofs for the Cantor-Bendixon theorem are evaluated. Although the proof via condensation points is simpler than the proof via the derivative, the derivative proof is of greater importance: two of the most standard textbooks of descriptive set theory, Kechris (1995) and Moschovakis (2009), point out that it is important for generalisations of the theorem, for example for analytic sets.[20] But they also consider this proof to be 'more informative' than the simpler proof via condensation points.[21]

We understand this use of 'more informative' at least partly to mean 'more explanatory': the more informative construction of the Cantor-Bendixon derivative provides us with greater insight into the general nature of these perfect sets. In other words, the easier construction (*P* as the set of condensation points) shows us what the perfect set looks like, but the construction via derivatives additionally shows us why the perfect set looks like that, and this holds not only for one theorem, but for all the dichotomy theorems before Silver's:

> One was therefore led naturally to the belief that the abundance of such derivatives is the driving force underlying the great variety of dichotomy theorems in descriptive set theory. (Miller, 2009)

What is this 'driving force' in terms of explanatoriness? Considering the proof using the iterated Cantor-Bendixon derivatives, its explanatoriness arises from the way in which the perfect set is built. We construct a set that consists solely of limit points by 'sorting out' more and more of the non-limit points and at the same time closing under limit points in transfinitely many steps. By construction, there is a point in these steps, α_0, where this process stabilises and naturally produces the desired set: the set that contains all and only its limit points. Following this construction, we are able to 'observe' how the perfect set grows out of the inherent properties of the construction (i.e., the definition of the derivative operation, the property that such an α_0 exists, etc.).

This construction fits a type of explanatoriness given by Steiner (1978).[22] Steiner identifies two components explanatory proofs should have, namely, they should refer to a characterising property of an entity in the theorem such that 'from the proof

[19] See Kechris (1995, 33–34) for the full proof.

[20] See, for example, Moschovakis (2009, 59–60) for Souslin's theorem

[21] See Kechris (1995, 33): 'a more informative construction of the kernel', the kernel being the perfect set; Moschovakis (2009, 59) calls the whole argument 'more informative'.

[22] Here we do not claim that Steiner's account of explanatoriness is the best or 'right' one; we see it as one way of giving an account for the explanatoriness of a proof. In our case, mathematicians see the proof as explanatory, so we use Steiner's theory to try to ascertain what features make this proof

it is evident that the result depends on the property' (Steiner, 1978, 143), and they should provide the possibility of generalising the feature(s) connected to this characterising property to produce other, related theorems and proofs. Both features are present in the classical proof type: the characterising property occurring in the theorem is the property of being perfect.[23] The classical proof refers to this property by showing how it can be produced through the derivative construction. So one fundamental feature of the property of being perfect is its construction via derivatives, and this feature fulfils Steiner's second criterion of generalisability. In the case of the early dichotomy theorems, the classical proof fulfils this in a very strong way, as varying this feature produces proofs for ever more general dichotomy theorems.

Colyvan et al. (2018) discuss the local dependence-based explanation, a more general form of explanation stemming from considerations in the philosophy of science.[24] Local dependence-based explanations involve constructions of mathematical objects that then exhibit the desired property in a deep way, meaning that the property '[does not only follow] logically … from the construction in question, rather, we mean that the … property naturally arises from the core properties of the construction in question' (Colyvan et al., 2018, 14). This fits the classical proof quite well; we already used terms like 'naturally produces' and 'grows out of' above to describe the construction of the perfect set.

We conclude that the classical proof for the early dichotomy theorems is explanatory, based primarily on the intuition of mathematicians such as Miller, Kechris, and Moschovakis. But it is interesting that this proof also fits well with a couple of philosophical accounts of mathematical explanation.

3.1.3 A 'New' Classical Proof

As we already noted, the classical proof type was not (and could not be) used any longer for Silver's theorem and later generalisations. However, around 2010 Ben Miller developed a new proof for the Kechris-Solecki-Todorcevic theorem,[25] a very advanced general dichotomy result. This exhibits features that are very similar to the classical proof type. Miller introduces what he calls the 'graph-theoretic approach to dichotomy theorems'.[26] This approach developed out of the modern work on dichotomy theorems such as Silver's that generalise the older theorems by focusing on definable equivalence relations, i.e., subsets of $X \times X$ for a Polish space X that are definable in a certain manner (e.g., analytic, Borel, etc.). Most notable are the new dichotomies introduced in Harrington et al. (1990), Hjorth and Kechris (1997), and Kechris et al. (1999), where the latter one specifically produces dichotomy theorems for graphs.[27]

explanatory. See Colyvan and Resnik (forthcoming) for discussion of Steiner's account and its influence in contemporary philosophy of mathematics.

[23] Some of these theorems are also called 'perfect set theorems'.

[24] See Colyvan et al. (2018, 13–16) for more information.

[25] See Kechris et al. (1999).

[26] See, for example, Miller (2012).

[27] A graph is an irreflexive symmetric subset of the product of the underlying Polish space.

Miller based his work specifically on the notions used in Kechris et al. (1999), using graphs and colourings of graphs to build up a new proof for the Kechris-Solecki-Todoercevic (KST) dichotomy theorem, which had until then only a proof of the advanced logic type. This proof is much too complex to present here; an excellent survey of this approach is presented in Miller (2012). We therefore refer to this exposition for more details and only discuss its details against the backdrop of the classical Cantor-Bendixon derivative proof given above.

Most importantly, the heart of Miller's proof consists of a transfinite construction similar to the iterative Cantor-Bendixon derivative construction, only here we transfinitely construct Borel sets on which a graph G has a Borel $\aleph_0$-colouring. Miller himself considers this construction to be the more complex analogue to the classical proof of the early dichotomy theorems: '[One] obtains a classical proof… resembling that of Cantor's perfect set theorem via the Cantor-Bendixson derivative' (Miller, 2012, 6).

Although presenting us with a much more complex situation than in the contexts of the earlier classical proofs, Miller's proof inherits its explanatory features from them. Miller himself remarks upon this fact in a lecture given on his work on the subject, when his proof was still work in progress (emphasis added):

> The new ideas described here appear to be leading towards a *classical explanation* of descriptive set-theoretic dichotomy theorems … the new proofs restore the intuition that the abundance of derivatives is at the heart of the matter. (Miller, 2009)

Returning to Steiner's account, we see a very similar picture to the classical proof of the early theorems: we have a construction that builds up the characterising feature we are after: the $\aleph_0$-colouring of a graph. The generalisability is a strong point in favour of this proof. Most of Miller (2012) is concerned with presenting various ways in which the graph-theoretic approach can be varied to produce a cornucopia of variations, specifications, and generalisations of the previously known dichotomy results. The generalisability of the graph-theoretic proof is therefore much greater than that of the standard classical proof, which ceased to be generalisable for Silver's theorem and later ones.

When looking at the local dependence-based model of explanation the situation is not that clear, as it will be nearly impossible for the average investigator into mathematical explanation to judge the 'naturalness' of the graph-theoretic approach to dichotomy theorems. The mathematical details are too complex and the mathematical theory on which it rests is too expansive to be clearly and deeply understood by more than a handful of experts in this area. We therefore have to rely on the intuition of the experts to find evidence for explanatoriness. In our case, this intuition points towards a strong analogy in the general set-up of the classical early proofs and the proof in the graph-theoretic setting, going so far as to see the different types of transfinite constructions used in the proof to produce the same type of entity, namely derivatives. Furthermore, both constructions are judged to be explanatory because of *the same way* in which they produce these derivatives (see Miller, 2009, and Miller, 2012). So based on our judgement that the classical proof is explanatory in the way in which it naturally gives rise to the property in question, the graph-theoretic approach does the same in its respective construction.

3.2 The Advanced Logic Proof

We now turn to the advanced logic proof type that was initially developed by Silver to prove a further generalisation of the early dichotomy theorems.

Theorem 3 (Silver). *If X is a Polish space and* $E \subseteq X^2$ *a* $\mathbf{\Phi}^1_1$ *equivalence relation, then either E has only countable many equivalence classes or there exists a perfect set of pairwise inequivalent elements.*[28]

The proof given by Silver (1980) and the improvements thereof by Harrington (1976) and Louveau (1979) are of a totally different kind than proofs by derivative-style constructions. The Silver proof makes use of an advanced logical set-up, relating to the other, perhaps more 'meta-mathematical' fields of set theory and mathematical logic in general.

Silver himself used three logical tools in the proof of his theorem, namely the technique of forcing—developed to show undecidability results in set theory—methods from effective descriptive set theory—a recursion theoretic analogue to descriptive set theory—and iterates of the Power Set axiom. Harrington simplified this by getting rid of the last technique (therefore he called it a 'powerless' proof in Harrington, 1976), but he still relied on forcing and methods from recursion theory (for a different version see also Harrington and Shelah, 1982). Finally, Louveau (1979) produced a proof that formulates the forcing part of Harrington's proof in a topological manner using the so-called Gandy-Harrington topology.[29] It therefore does not use forcing, but still relies on techniques from topology and effective descriptive set theory.[30]

[28] The expressions Σ^n, Φ^n, Δ^n mark the complexity of a set (or formula that defines a set) with respect to some mathematical concept. According to this complexity, hierarchies of sets can be given. One example is the Borel hierarchy, where the complexity is measured with respect to taking countable unions and complements; a different example is the Lévy hierarchy, where formulas are more complex the more often their unbounded quantifiers change. Here Φ^1 refers to a coanalytic set, meaning that it is a complement of an analytic (Σ^1) set; it is part of the Projective hierarchy.

[29] For a definition see Harrington et al. (1990, 917).

[30] Such a topological version of Harrington's proof is also given in Martin and Kechris (1980), however, according to Harrington et al. (1990, 907) it is based on seminar notes of Louveau.

Both the proof of Harrington[31] and the proof type of Louveau are still in use, however the topological (and not forcing-related) approach of the latter seems to be used more often in the proof of recent dichotomy results. In particular, the use of effective methods is more essential than the use of forcing because it is used in both proof types and because the relationship between effective descriptive set theory (EDST) and the non-effective, classical descriptive set theory (CDST) has more wide-ranging applications, unrelated to dichotomy theorems. We will therefore focus our account of explanatoriness on this advanced logic part of the proof type.

Effective descriptive set theory goes back to the work of Kleene.[32] Initially he developed it outside of descriptive set theory by using methods from recursion theory to study sets. Instead of considering sets that are definable in a certain manner, one studies their recursion-theoretic analogues (that still are definable in a certain manner). As a basic example, a set is recursive if it is computable in the sense that there exists an algorithm that always decides in a finite amount of time whether something is an element of the set or not. Likewise, a recursively enumerable set is one in which the algorithm decides in the above way whether something is an element, but can sometimes return no answer. These notions can be spelled out mathematically via functions and are basic notions in the field of recursion theory.

It turns out that there is a whole field of research analogous to CDST that makes use of these recursion-theoretic notions. As an example, let us consider basic sets studied in CDST. Here a $\boldsymbol{\Delta}^{\mathbf{0}}$ pointclass (in bold font) is one where the elements are closed and open. It corresponds to the (non-bold font) Δ^0 pointclass that consists of the recursive pointsets; likewise the $\boldsymbol{\Sigma}^{\mathbf{0}}$ pointclass (open sets) correspond to the Σ^0 pointclass (recursively enumerable sets) and so on (the effective versions are the parameter-free versions of the classical definitions). Based on these relations, whole hierarchies of sets can be built up in CDST that have an analogous version in EDST, giving rise to theorems that have classical versions (the CDST version) and effective versions in the formulation of EDST.

Based on Kleene's work, Addison developed the exact analogies between CDST and EDST. Since then, mathematical work has shown that this is not a local phenomenon but holds on a fundamental level in many areas of DST.[33]

It therefore turns out that CDST and EDST are very tightly interconnected up to a point where both can be seen to be refinements of the other (emphasis in the original):

> Over the years and with the work of many people, what was first conceived as 'analogies' developed into a general theory which yields in a unified manner both

[31] See, for example, Miller (1995, 113–115).

[32] For short introductions into EDST with reference to dichotomy theorems, see Harrington et al. (1990, Ch. 3) or Martin and Kechris (1980). For a more general account, see, for example, Moschovakis (2009).

[33] For more on the historical development, see Kanamori (1995) and Moschovakis (2009, introduction).

> the classical results and the theorems of the recursion theorists; more precisely, this effective theory yields *refinements* of the classical results. (Moschovakis, 2009, 5)[34]

We will now study what role this relation plays in the proof of dichotomy theorems; in particular we will study the example of the KST theorem (Theorem 6.3 in Kechris et al., 1999), as this was also the theorem for which the first 'new' classical proof was developed.

Theorem 4 (Kechris-Solecki-Todorcevic). *Let X be a Polish space and* $\mathcal{G}(X,R)$ *an analytic graph (i.e.,* $R \subseteq X^2$ *is analytic). Then exactly one of the following holds:*

(1) $\chi_B(\mathcal{G}) \leq \aleph_0$

(2) $\mathcal{G}_0 \leq_c \mathcal{G}$

Here $\mathcal{G}_0$ is a certain minimal graph and $\chi_B(\mathcal{G})$ the Borel chromatic number of the graph (all of the definitions can be found in Kechris et al., 1999). We don't need to understand the exact definitions of the notions involved here to study this as an example of the set-up of the advanced logic proof type. For that let us look at how the authors of the theorems begin its proof. Directly after stating the theorem, they continue in the following manner:

> This result is proved using methods of effective descriptive set theory, in particular the Gandy-Harrington topology. In fact one has the following effective version (which by standard arguments implies the above theorem). (Kechris et al., 1999, 21)

They proceed to give the effective version:

Theorem 5. *Let* $\mathcal{G} = (\mathbb{N}^{\mathbb{N}}, R)$ *be a* Σ_1^1*-graph (i.e.,* $R \subseteq (\mathbb{N}^{\mathbb{N}})^2$ *is* Σ_1^1*). Then exactly one of the following holds:*

(1) *There is a* Δ_1^1 *colouring c:* $\mathbb{N}^{\mathbb{N}} \to \mathbb{N}$ *for* $\mathcal{G}$;

(2) $\mathcal{G}_0 \leq_c \mathcal{G}$

Again, we don't need to grasp all the relevant definitions to see the main point of this approach: we have a theorem that can be presented in two versions, one using classical notions and one using effective notions from descriptive set theory (e.g., in the first, the graph is given as analytic, which is boldface $\boldsymbol{\Sigma}^1$; in the second, it is

[34] For the other direction of refinements, see Moschovakis (2010).

lightface Σ^1, the effective version). This procedure goes back to proofs of dichotomy theorems in Harrington et al. (1990), where the classical results are obtained by relativising the effective version to a parameter.[35]

To summarise the above: the advanced logic proof does not show the connections with derivatives; in fact, it represents a clear break with the proofs of the earlier dichotomy theorems. Instead, it links the dichotomy theorems to EDST as well as providing a further example for the connection of CDST and EDST. Thereby it both uses and strengthens the interconnection between CDST and its effective counterpart. Through this interconnection, the proof situates the dichotomy theorems in the larger context of this general feature of descriptive set theory and therefore unifies dichotomy theorems with other results that make use of this feature as well.

This interconnection also lies at the heart of the explanatory value that advanced logic proof provides us with. We mentioned above that the classical proof type fits the local dependency-based account of explanation as given in Colyvan et al. (2018). There the authors also outline another kind of explanation, the global unification-based explanation:[36]

> [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. (Colyvan et al., 2018, 15)

As we are already looking at a certain class of theorems—the dichotomy theorems in DST—that is defined by the common characteristics of its members, let us rephrase this for this situation: a class of theorems is explained by deriving its members using a proof type that shows that the class of theorems is part of a very general, perhaps utterly pervasive, pattern that is characteristic for the area of mathematics it is a part of.

The advanced logic proof type provides us with this kind of explanatory value: looking at the pervasive pattern of the close analogies between CDST and EDST that have shown themselves in various areas before the advanced logic proof shows how dichotomy theorems are a very fruitful part of this pattern. As one example, consider how this back and forth between DST and its effective version is used in the proofs of dichotomy theorems such as Theorem 6.3 and Theorem 6.4 in Kechris et al. (1999, 21).

[35] See Harrington et al. (1990, 916) for an introduction to such relativisations; the complete proof on which the proof of the KST theorem depends on in a crucial manner can be found in Harrington et al. (1990, 919–927).

[36] This account can also be related to Kitcher's unificatory account for explanations in mathematics and the sciences; see, for example, Kitcher (1989). Indeed we think a case can be made that the advanced logic type is a very good candidate for what Kitcher calls the explanatory store for a system of beliefs.

This back and forth can also be applied to earlier dichotomy theorems that are usually proven via the classical proof and give rise to concrete examples of explanation. For an example that provides a concrete way in which the relation to EDST can be explanatory, consider the following:[37]

Theorem 6 (Souslin, 1917). *Every uncountable analytic set has a non-empty perfect subset.*

Moschovakis points out[38] that an effective version of this result provides an explanation of the theorem 'in terms of definability rather than size: an analytic set P has a perfect subset if it has at least one member which is *more difficult to define* than P itself.' Such an effective version is due to a result by Harrison from 1967 (see Harrison 1967). After Silver's theorem provided a proof of the advanced logic type for the first time, it thereby established the connections to effective DST. So although Souslin's theorem can be proven directly by an easier classical-type proof, the connection to effective DST provides a different kind of explanation.

This is also emphasised by Ramez Sami, who points out the added value the advanced logic proofs provide:

> The present note greatly antedates the more recent 'back-to-classical' movement developed *con maestria* [by Ben Miller]. We still hold that effective methods will often yield simpler proofs of stronger and finer results. (Sami, 2019, 4039)

A similar sentiment with regard to explanatoriness was expressed by Sami in an online event with one of the authors on 25 November 2020. We therefore have a similar situation for the advanced logic proof type as with the classical proof type: intuitions by mathematicians tell us that the proof is explanatory, and we can back this up by showing how it fits with at least some of the philosophical accounts of mathematical explanation.

4 Conclusions

The dichotomy theorems in DST are good examples of how questions of explanatoriness not only play a role in, but also can direct, mathematical research.[39] The lack of a classical proof for Silver's theorem and further generalisations was the main motivation for searching, and finally producing, a different

[37] The authors would like to thank Yiannis Moschovakis for pointing out this example.
[38] In private communication with the authors; e-mail from 12 September 2020.
[39] This has been called into question, for example by Zelcer (2013).

proof that supplies us with a deeper understanding of the theorem, showing not only that it holds but also *why* it holds.[40] This new proof led to a reconsideration of the field of dichotomy results in descriptive set theory, not only by reproving already known theorems, but also by introducing a new approach to this area that produces new theorems and new generalisations.

These arguments and others given in Section 3.1 show that the classical proof can be considered to be explanatory. However, we have also seen that there are arguments for the explanatoriness of the advanced logic proof. This points towards a pluralist picture of what explanation is in mathematics. Taking the intuitions of mathematicians as our evidence, it might be argued that one type of explanatoriness is as valid as the other, precisely because mathematicians' intuitions do not converge.

If we accept such a pluralist picture of explanatoriness, we might ask where this pluralism comes from. For the case of dichotomy theorems, let us outline one possibility.[41] Here, we can relate the different intuitions about explanatoriness to the communities of different sub-areas of DST. So, for researchers who primarily consider dichotomy theorems from the classical point of view, focusing on inherent similarities of the dichotomy theorems 'from within' (i.e., the way in which they are built up via derivatives and the inherent properties of objects like perfect sets), the classical proof is more explanatory because it provides one with a vivid picture of how the inherent properties of the objects in question give rise to the various theorems.

If one approaches the dichotomy theorems from the viewpoint of general descriptive set theory, where a lot of research has gone into the interconnections between CDST and EDST, we consider the theorems 'from outside' and see them as one example for a more general pattern in the theory. The advanced logic proof is thus seen as more explanatory because it ties in with these general patterns.

Indeed, such sentiments were expressed by the mathematicians themselves. B. Miller, for example, mentioned that people who work deeply with dichotomy theorems prefer the classical proof, while it might be different for people whose work is more closely connected to effective DST.[42] So what mathematicians see as explanatory seems to be motivated by the epistemic interest they pursue in their practice. These epistemic interests can be the search for basic entities that lie at the heart of a collection of theorems, the search for overarching patterns in mathematical reasoning, etc.

[40] B. Miller confirmed that this was his main motivation in reproving Silver's theorem and generalisations thereof. (Personal communication with one of the authors; virtual meeting on 8 July 2020)

[41] We do not claim that this is the only or even the main reason for every case of pluralism in explanations. There is also the difficult task of distinguishing explanatoriness from other virtues found in good mathematics and, indeed, determining what makes for good mathematical proofs—such as those that Paul Erdös referred to as coming from God's book of the best proofs or simply 'from the book' (Aigner and Ziegler, 2010). The virtues of good mathematics include beauty, simplicity, elegance, explanatory power, and so on (see Tao, 2007; Inglis and Aberdein, 2015).

[42] Personal communication with one of the authors; virtual meeting on 8 July 2020.

We conclude that our main case study—the dichotomy theorems in descriptive set theory—as well as the elementary example we started with—Fermat's Little Theorem from elementary number theory—suggest a kind of explanatory pluralism. It would appear that whether a proof is seen as more explanatory than another depends on the epistemic interests and goals of the practitioners. These epistemic interests can be connected to (unofficial) research agendas or other practices of sub-communities of a discipline. In this sense, the explanatoriness or otherwise of a proof must be assessed in relation to the wider mathematical context in which it sits. Perhaps this is not surprising. But accepting such a context-sensitive notion of explanation in mathematics would be a serious blow for those of us with monist leanings—those who seek a single, elegant, and unified account of mathematical explanation. For better or worse, we have good reason to believe that explanation in mathematics is more complex and more interesting than the monist would wish.

This raises interesting questions about the nature of any potential pluralism. For instance, if there is explanatory pluralism in mathematics, does this arise because there are different levels of explanation in operation in mathematics? As we have seen, there are different levels of abstraction in mathematics and explanations arising at these different levels. It seems a good working hypothesis that these different explanations correspond to different levels of explanation. But an alternative might be that the different explanations are in fact answering different why questions—perhaps why questions pitched at different levels of abstraction. What might these different why questions be? In our experience, mathematicians tend to be interested in explanation in an apparently unitary sense: *why does the theorem in question hold*. On the face of it, at least, this looks like a single why question, but appearances may be deceptive here. Indeed, the lack of precision in the question 'why does the theorem hold?' may be hiding ambiguity about exactly what is being asked. Is it an ambiguous question or is it inviting different levels of explanation?

These interesting issues require much further work. Any attempt to settle them now would be misguided. As things currently stand, philosophical work on mathematical explanation is in its infancy. In our view, work on mathematical explanation requires more case studies to draw upon before we tackle some of the questions just raised about the nature of any potential explanatory pluralism in mathematics.[43] The purpose of this paper is to provide a couple more case studies that we trust will be helpful in addressing some of the many puzzles about mathematical explanation.[44]

[43] For obvious reasons, we do not want to develop a philosophical account of mathematical explanation based on a few examples from one or two areas on mathematics. Just as an account of scientific explanation needs to work across all areas of empirical science, an account of mathematical explanation should work across all areas of mathematics.

[44] We'd like to thank Neil Barton, Hazel Brickhill, Erik Curiel, Clio Cresswell, Ben Eva, Stephan Hartmann, Anthony Henderson, Leon Horsten, Silvia Jonas, Deborah Kant, Daniel Kuby, Randall

References

M. Aigner and G. M. Ziegler. *Proofs from the Book*. Springer, Heidelberg, 4th edition, 2010.

S. Baron, M. Colyvan, and D. Ripley. A counterfactual approach to explanation in mathematics. *Philosophia Mathematica*, 28(1): 1–34, 2020.

G. Cantor. Über unendliche, lineare Punktmannigfaltigkeiten. *VI, Mathematische Annalen*, 23: 210–246, 1884.

M. Colyvan. *An Introduction to the Philosophy of Mathematics*. Cambridge University Press, Cambridge, 2012.

M. Colyvan. The ins and outs of mathematical explanation. *Mathematical Intelligencer*, 40(4): 26–29, 2018.

M. Colyvan and M. D. Resnik. Explanation and realism: Interwoven themes in the philosophy of mathematics. In Y. ben Menahem and C. Posy, editors, *Mathematical Objects, Knowledge and Applications: Essays in Memory of Mark Steiner*. Springer, New York, forthcoming.

M. Colyvan, J. Cusbert, and K. McQueen. Two flavours of mathematical explanation. In A. Reutlinger and J. Saatsi, editors, *Explanation Beyond Causation: Philosophical Perspectives on Non-Causal Explanations*, pp. 231–249. Oxford University Press, Oxford, 2018.

W. D'Alessandro. Mathematical explanation beyond explanatory proof. *British Journal for the Philosophy of Science*, 71(2): 581–603, 2020.

L. Euler. Theoremata circa residua ex divisione potestatum relicta. *Novi Commentarii Academiae Scientiarum Petropolitanae*, 7: 49–82, 1761.

H. Field. *Science Without Numbers: A Defence of Nominalism*. Oxford University Press, Oxford, 2nd edition, 2016.

S. W. Goloumb. Combinatorial proof of Fermat's "little" theorem. *American Mathematical Monthly*, 63(10): 718, 1956.

J. Hafner and P. Mancosu. The varieties of mathematical explanation. In P. Mancosu, K. F. Jørgensen, and S. A. Pedersen, editors, *Visualization, Explanation and Reasoning Styles in Mathematics*, pp. 215–250. Springer, Dordrecht, 2005.

L. Harrington. A powerless proof of a theorem by Silver. *Handwritten note dated 11–76*, 1976.

L. Harrington and S. Shelah. Counting equivalence classes for co-κ-Souslin equivalence relations. In D. Van Dalen, D. Lascar, and T. Smiley, editors, *Logic Colloquium '80*, volume 108 of *Studies in Logic and the Foundations of Mathematics*, pp. 147–152. Elsevier, Amsterdam, 1982.

L. A. Harrington, A. S. Kechris, and A. Louveau. A Glimm-Effros dichotomy for Borel equivalence relations. *Journal of the American Mathematical Society*, 3(4): 903–928, 1990.

J. Harrison. *Some Applications of Recursive Pseudo-Well Orderings*. PhD thesis, Stanford University, Stanford, 1967.

G. Hjorth and A. S. Kechris. New dichotomies for Borel equivalence relations. *The Bulletin of Symbolic Logic*, 3(3): 329–346, 1997.

M. Inglis and A. Aberdein. Beauty is not simplicity: an analysis of mathematicians' proof appraisals. *Philosophia Mathematica*, 23(1): 87–109, 2015.

A. Kanamori. The emergence of descriptive set theory. In J. Hintikka, editor, *From Dedekind to Gödel: Essays on the Development of the Foundations of Mathematics*, pp. 241–262. Springer Netherlands, Dordrecht, 1995.

McCutcheon, Ben Miller, Yiannis Moschovakis, Hannes Leitgeb, Miklós, Rédei, Máté Wierdl, Geordie Williamson, and Alistair Windsor for very helpful discussions on the topic of this paper. We are especially indebted to Alastair Wilson for detailed comments on an earlier draft of this paper and for his many helpful suggestions. Carolin Antos's work on this paper was supported by the Volkswagen Foundation (Freigeist Grant) and the Zukunftskolleg (University of Konstanz, Germany). Mark Colyvan's work on this paper was supported by an Australian Research Council Future Fellowship (grant number: FT110100909) and a Carl Friedrich von Siemens Research Award of the Alexander von Humboldt Foundation.

A. Kechris. *Classical Descriptive Set Theory*, volume 156 of *Graduate Texts in Mathematics*. Springer-Verlag New York, 1995.

A. Kechris, S. Solecki, and S. Todorcevic. Borel chromatic numbers. *Advances in Mathematics*, 141(1): 1–44, 1999.

P. Kitcher. Explanatory unification and the causal structure of the world. In P. Kitcher and W. C. Salmon, editors, *Scientific Explanations*, pp. 410–505. University of Minnesota Press, Minneapolis, 1989.

M. Lange. *Because Without Cause: Non-Causal Explanations in Science and Mathematics*. Oxford University Press, Oxford, 2017.

M. Lange. Mathematical explanations that are not proofs. *Erkenntnis*, 83(6): 1285–1302, 2018.

D. K. Lewis. Causal explanation. In Philosophical Papers Vol. 2, pp. 214–240. Oxford University Press, Oxford, 1986.

A. Louveau. Une nouvelle technique d'étude des relations d'équivalence coanalytiques. In S. Grigorieff, K. McAloon, and J. Stern, editors, *Set Theory: GMS Seminar (Paris, 1976–1977 and 1977–1978)*, volume 5 of *Publ. Math. Univ. Paris VII*, pp. 35–42. Univ. Paris VII, Paris, 1979.

P. Mancosu. Mathematical explanation: why it matters. In P. Mancosu, editor, *The Philosophy of Mathematical Practice*, pp. 134–149. Oxford University Press, Oxford, 2008.

D. A. Martin and A. Kechris. Infinite games and effective descriptive set theory. In C. A. Rogers, J. E. Jayne, C. Dellacherie, F. Topsoe, J. Hoffman-Jorgensen, D. A. Martin, A. S. Kechris, and A. H. Stone, editors, *Analytic Sets*, pp. 404–470. Academic Press, San Diego, 1980.

A. W. Miller. *Descriptive Set Theory and Forcing: How to Prove Theorems about Borel Sets the Hard Way*. Lecture Notes in Logic. Springer-Verlag Berlin, Heidelberg, 1995.

B. D. Miller. *Forceless, ineffective, powerless proofs of descriptive set-theoretic dichotomy theorems. Slides of talk held at Logic Colloquium*, 2009. URL http://lc2009.fmi.uni-sofia.bg/presentations/bmiller.pdf.

B. D. Miller. Dichotomy theorems for countably infinite dimensional analytic hypergraphs. *Annals of Pure and Applied Logic*, 162(7): 561–565, 2011.

B. D. Miller. The graph-theoretic approach to descriptive set theory. *Bulletin of Symbolic Logic*, 18(4): 554–575, 2012.

Y. Moschovakis. *Descriptive Set Theory*, volume 155 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence RI, 2nd edition, 2009.

Y. N. Moschovakis. Classical descriptive set theory as a refinement of effective descriptive set theory. *Annals of Pure and Applied Logic*, 162(3): 243–255, 2010.

A. Reutlinger, M. Colyvan, and K. Krzyz˙anowska. Prospects for a monist theory of non-causal explanation in science and mathematics. *Erkenntnis*, 87(4): 1773–1793, 2022.

K. H. Rosen. *Elementary Number Theory and Its Applications*. Addison-Wesley, Redwood City, CA, second edition, 1988.

R. L. Sami. A note on an effective Polish topology and Silver's dichotomy theorem. *Proceedings of the American Mathematical Society*, 147(9): 4039–4044, 2019.

J. H. Silver. Counting the number of equivalence classes of Borel and coanalytic equivalence relations. *Annals of Mathematical Logic*, 18(1): 1–28, 1980.

M. Souslin. Sur une definition des ensembles mesurables B sans nombres transfinis. *Comtes Rendus de l'Acad ́emie des Sciences: Series 1 Mathematics*, 164: 88–91, 1917.

M. Steiner. Mathematical explanation. *Philosophical Studies*, 34(2): 135–151, 1978.

T. Tao. What is good mathematics? *Bulletin of the American Mathematical Society*, 44(4): 623–634, 2007.

A. Weil and M. Rosenlicht. *Number Theory for Beginners*. Springer, New York, 1979.

M. Zelcer. Against mathematical explanation. *Journal for General Philosophy of Science*, 44(1): 173–192, 2013.

14
A Dormitive Virtue Puzzle

Elanor Taylor

BACHELIERUS
Mihi a docto doctore
Domandatur causam et rationem quare
Opium facit dormire.
A quoi respondeo,
Quia est in eo
Vertus dormitiva,
Cujus eat natura
Sensus assoupire.

I am asked by the learned doctor for the cause and reason that opium makes one sleep.

To this I reply that there is a dormitive virtue in it, whose nature it is to make the senses drowsy.

CHORUS
Bene, bene, bene, bene respondere.
Dignus, dignus est intrare
In nostro docto corpore.

Very, very, well answered. The worthy [candidate] deserved to join our learned body.[1]

1 Introduction

In a rambunctious scene in Molière's 1673 comedy *The Imaginary Invalid*, the performers pantomime a medical student's qualifying examination. At a key comedic

[1] Latin quotes from Molière, Jean Baptiste (1926) Vol. 8, pg 328. English translation given by Hutchison in Hutchison, Keith (1991) pg 245

Elanor Taylor, *A Dormitive Virtue Puzzle* In: *Levels of Explanation.* Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Elanor Taylor 2024. DOI: 10.1093/oso/9780192862945.003.0015

moment, the doctors ask the student to explain why opium induces sleep, and the student replies that opium has a "dormitive virtue." The doctors break into applause and admit the candidate to the medical profession.[2] Molière framed this scene as a satirical play on the use of opaque scholastic concepts in medicine, and since then the phrase "dormitive virtue" has become a byword for explanatory failure.

Opinions differ as to why the dormitive virtue explanation is so bad.[3] As we will see, some cite its apparent circularity, others its mysteriousness, and others its lack of causal detail. From the early modern period to the present day, however, a majority of philosophers agree that the dormitive virtue explanation fails and that if a philosophical position permits explanations like this, then that is a prima facie count against the view. However, as I will discuss in Section 3, contemporary work on the metaphysics of grounding and dispositions appears to permit explanations in which dispositions explain patterns in events, with a structure strikingly similar to the dormitive virtue case. Furthermore, if we consider this case away from its early modern comedic context we may find ourselves wondering what is so bad about this attempt at explanation. Does it not give us *something* useful in pointing to the opium rather than to the sleepers, or to their surroundings? Does it not at least tell us *where* to look for an explanation? Anyone who takes the pragmatic aspects of explanation seriously is likely to find such considerations familiar and compelling. Viewed from these different perspectives—metaphysics on the one hand, and pragmatist philosophy of science on the other—the dormitive virtue case does not look so bad after all.

There are good historical reasons for conflicting responses to this case. According to some interpretations many prominent thinkers of Molière's period defined themselves in opposition to the scholastic era, rejecting the apparatus of medieval metaphysics that included reified powers. Contemporary metaphysics, on the other hand, has embraced the neo-Aristotelian resources central to medieval thought.[4] However, I am interested in focusing on this case independently of these issues about historical interpretation. Examined in isolation from its history the dormitive virtue explanation presents a puzzle: one the one hand it appears to be an obvious explanatory failure, while on the other it looks like a perfectly adequate explanation.

In this paper I will address this puzzle. My first goal is to articulate and motivate the puzzle, in making more precise the nature of dormitive virtue explanation, the considerations driving positive and negative verdicts on this case, and what is at stake in searching for a resolution. Then I sketch a view of explanation that

[2] The "medical student" in this scene is the hypochondriac of the play's title, making the pantomime examination even more absurd.

[3] For ease of expression I will refer to the "dormitive virtue explanation" and "dormitive virtue case" as coextensive, although part of what is at issue is whether or not this case involves a genuine explanation. I will also take "virtue," "power," "faculty," and "disposition" to be coextensive.

[4] For example, see Tahko, Tuomas, E. (ed.) (2011).

illuminates the puzzle, shows what is required for a resolution, and makes sense of conflicting responses to this case. I call this approach Contextualist Pluralist Non-Realist Backing, or CPN Backing. Showing that this view offers useful resources for addressing the dormitive virtue puzzle will not constitute an argument for CPN Backing, but it will illustrate some of its attractive features as a model of explanation. In particular, I will show that CPN Backing is unusual in taking connections between explanation and metaphysics seriously while also prioritizing contextual and pragmatic aspects of explanation, and that this combination offers a helpful approach to the dormitive virtue case. The end result will be an improved understanding of dormitive virtue explanation and of how to resolve the dormitive virtue puzzle, and an illustration of the advantageous features of CPN Backing as an approach to explanation.

2 Dormitive virtue explanation and its discontents

Those who raise concerns about dormitive virtue explanations tend to focus on this version:

> DV: opium reliably induces sleep because it has a dormitive virtue.

This can be generalized to other cases:

> DV General: F reliably φs because F has a φ-ing virtue.

There are other explanations of apparently similar structure that cite virtues. For example:

> L: Laura fell asleep because she ingested opium, which has a dormitive virtue.

Alternatively:

> LS: Laura fell asleep more quickly than Sarah because Laura ingested opium, which has a dormitive virtue, while Sarah did not.

However, most controversy about the dormitive virtue explanation focuses on DV and explanation with the structure DV General, rather than individual or contrastive cases like L and LS. This is perhaps because DV and DV General are comparatively less plausible as explanations, while L and LS are more acceptable. For example, Vetter holds that dispositions can play a central role in contrastive explanations like LS, while some causal theories of explanation permit L as at least a partial causal explanation.[5] And, although faithfulness to Molière's words is typically not prioritized in these discussions, DV roughly captures the explanation offered in the original play.

In Molière's time many authors were troubled by the prospect of DV and explanations of the form DV General. For example, Glanvill discussed the claim that fire burns in virtue of its heat and described it as "an empty dry return to the Question," and, "no better account than we might expect of a Rustick."[6] Locke raised a similar worry about the explanatory pointlessness of faculties when he wrote, "For *faculty, ability,* and *power*, I think, are but different names of the same things: which ways of speaking, when put into more intelligible words, will, I think, amount to thus much: That digestion is performed by something that is able to digest, motion by something able to move, and understanding by something able to understand."[7] Malebranche was concerned about a tendency he observed among philosophers to, on encountering some new effect, posit an entity responsible for that effect. As he put it, "Fire heats things—therefore there is something in fire that produces this effect, something different from the matter of which fire is composed. And because fire is capable of several different effects (such as disintegrating bodies ... drying them, hardening them, softening them, enlarging them ... and so on), they liberally bestow on fire as many faculties or real qualities as effects it is capable of producing."[8] A number of other authors from this time, including Leibniz and Newton, raised concerns about explanations of this form.[9]

Many contemporary authors have also raised worries about the prospect of dormitive virtue explanation. For example, when discussing the commitments of the simple realist, Thomasson argues, "Not only does the simple realist not need to appeal to explanatory power or the like to justify her acceptance of the relevant entities, she cannot do so. Any attempt to do so would yield only a dormitive virtue explanation."[10] When considering the explanatory role of symmetry considerations, French notes, "One could of course suggest that white dwarf stars have a disposition to behave in such a way under gravitational collapse but that sails awfully close to a 'dormitive virtue' scenario."[11] While discussing evolutionary explanations that appeal to the notion of fitness, Sober argues that fitness cannot

[5] See discussion in Vetter, Barbara (2015) pg 87–89. Causal accounts of explanation that permit L as at least partial explanation include those defended in Lewis, David (1986) and Skow, Bradford (2014). An interventionist approach might also admit L in so far as the opium functions as a difference-maker. See Woodward, James (2003).

[6] Glanvill, Joseph (1665) pg 143. This section is discussed by Ott in Ott, Walter (2009) pg 11 ff.

[7] Locke, John (1689) 2.21.20.

[8] Malebranche, Nicolas, translated by Olscamp, Paul J. (1997) pg 242. This section is discussed by Hutchison in Hutchison, Keith (1991).

[9] See discussion in Hutchison, Keith (1991).

[10] Thomasson, Amie (2014) pg 156.

[11] French, Steven (2019) pg 26.

explain certain outcomes (even if it causes those outcomes) because this would amount to a dormitive virtue explanation.[12] In these cases we can see authors using the prospect of dormitive virtue explanation as a count against a view, such that if a philosophical strategy permits DV or DV General explanation, then that strategy must be abandoned. More directly engaging with the case, Strevens points out that his kairetic account of explanation does not permit dormitive-virtue-style explanations because they do not display sufficient causal detail, or "depth," for explanation.[13] McKitrick, when discussing the prospect of bare dispositions, unfavorably compares DV with an explanation that gives information about the causal mechanism through which opium induces sleep.[14]

As we can see from these extracts, there are a number of distinct concerns about DV and DV General explanations.

An initial worry is that the dormitive virtue explanation displays insufficient "explanatory distance."[15] Canonically explanation is irreflexive, so whatever the relata of explanation are—propositions, facts, sentences—they must be distinct. On this line of thought the dormitive virtue case fails to meet this standard because the fact that "opium reliably induces sleep" is the same as the fact that "opium has a dormitive virtue," in that the dormitive virtue is nothing beyond the pattern in events. This complaint about DV lies at the heart of many historical and contemporary concerns about it. For example, the idea that explanation by virtue or faculty is pointless, or as Glanvill put it, a "return to the question," indicates that the problem with the attempt at explanation is its circularity.[16] Furthermore, this worry about distance makes sense specifically of empiricist concerns about this case because a certain kind of empiricist can only countenance virtues as patterns in events, rather than as unobservable entities posited to explain such patterns.

Let us imagine that the proponent of the dormitive virtue explanation pushes back against this concern about irreflexivity. They argue that the explanation is not circular because the dormitive virtue is not merely a pattern in events but is instead a distinct entity, a virtue, that explains those events and is responsible for them. The circularity worry is straightforwardly avoided because the fact that opium has a dormitive virtue is distinct from the fact that opium reliably induces sleep. However, this leads us to the next complaint about the dormitive virtue explanation, which is that the entity it posits and the resulting explanation are *mysterious*. Following this line of thought, Goodman included dispositions on a list including "counterfactual assertions … angels, devils and classes" as entities that are "inacceptable without explanation."[17] In the early modern period the term

[12] Sober, Elliott (1984) pg 77–78.
[13] Strevens, Michael (2008) pg 131–133.
[14] McKitrick, Jennifer (2003) pg 349.
[15] The author coins this phrase and discusses this concern about DV in Taylor, Elanor (2023).
[16] Glanvill, Joseph (1665) pg 143.
[17] Goodman, Nelson (1983) pg 33.

"occult power" was used pejoratively to express the idea that such powers are unobservable, that their nature is inscrutable, and that there is no empirical basis for belief in their existence beyond the patterns in events they are posited to explain.[18] If powers are mysterious, then not only are they metaphysically troubling, but the purported explanation they support also does not provide the understanding or illumination we might expect from explanation. These two objections work together as a dove-tailing package: either the dormitive virtue explanation is circular, or else it is worryingly mysterious.

A third set of worries about explanations of this kind are generated by concerns about causation. There are a number of different objections here, but I will group them together as causal problems, oriented around the idea that DV displays some deviant connection between the dormitive virtue, the explanation, and the causal information relevant to the explanandum.

One such worry is that virtues are causally, and hence explanatorily, excluded by their categorical bases. On this line of thought DV is not a genuine explanation because it does not give us information about the real explanatory action which takes place in the categorical base of the dormitive virtue.[19] The dormitive virtue explanation is a placeholder for the causal-mechanical detail that explains the pattern in events, and renders talk of the dormitive virtue explanatorily redundant. However, we need not endorse such a strong exclusion principle to think that these explanations do not target the right level of causal detail. For example, as mentioned earlier, Strevens discusses DV and holds that explanations of this form lack the requisite causal detail to explain because they are pitched at too high a level of abstraction, rather than because they are excluded by an explanation given in categorical terms.[20] The idea that there is an appropriate level for causal explanation, whether of detail, of abstraction, or of scientific theory, has been taken up in conversation about levels of explanation across contemporary philosophy of science and metaphysics, including in Parts II and III of this volume.

Other causal considerations arise in the literature on dispositions. For instance, in response to the view that dispositions neither cause nor explain their instances, Sober and Shapiro argue that even if a disposition does not explain its instances it may still cause them, because they reject the background view linking explanation to causation and so can countenance a cause that does not explain.[21] But all parties to the conversation agree that, regardless of what is going on causally, DV General explanation fails.

[18] See discussion in Ott, Walter (2009) Ch. 5.

[19] This line of thought is evident in Prior, Elizabeth W., Pargetter, Robert, and Jackson, Frank (1982) pg 255.

[20] Strevens, Michael (2008) pg 131–133.

[21] Shapiro, Larry, and Sober, Elliott (2007) pg 19. Lange also discusses this extract in Lange, Marc (2017) pg 425.

A final concern applies to DV but not necessarily to all explanations of the form DV General. This is the worry that the dormitive virtue is too specific, coarse, or non-fundamental to play a central role in explanation, though other powers may play such roles. Some historical commentators have attributed this concern to Newton, who permitted some powers to play fundamental explanatory roles, but not powers as specific as the dormitive virtue.[22] This is an interesting worry because it hones in on something specifically wrong with the dormitive virtue case rather than something amiss with all DV General explanations. However, for this reason this is probably the most peripheral concern about the dormitive virtue case. Negative verdicts on the dormitive virtue case tend to be driven by more general, and generalizable, concerns about DV General explanation.

These four sets of objections may not exhaust the history of complaints about dormitive virtue explanation. Nor are they exclusive; for instance, a version of the causal exclusion worry can be motivated by considerations about explanatory distance. But overall, considerations about distance, mystery, causation, and specificity drive most of the historical and contemporary objections to DV and DV General explanation.

3 Contemporary dormitive virtue explanation

In Section 2 we saw the negative side: a range of commentators raising concerns about DV and DV General and arguing that if a view permits DV General explanation, then that is a count against the view. In this section we will see the positive side: contemporary work in metaphysics, particularly on grounding and dispositions, that appears to permit DV General explanation.

First, however, a note on pragmatism. Most pragmatists about explanation resist offering general accounts of explanation, beyond schematic ideas such as Van Fraassen's claim that explanations offer answers to why-questions, or Achinstein's view that explaining is a certain kind of speech-act.[23] Pragmatists will often warn against asking whether an explanation is legitimate in general, holding that it is only within a particular context that such questions can be meaningfully addressed. Accordingly, a pragmatist may have no difficulty with the idea that a certain explanation is both good and bad, and will simply say that it is good in some contexts and bad in others. On this approach the mere fact that the dormitive virtue explanation appears to be both bad and good does not raise any challenge or puzzle.

[22] For example, see discussion in Hutchison, Keith (1991).

[23] In Achinstein, Peter (1983), Van Fraassen, Bas (1980). For a discussion of concerns about general theories of explanation, see Díez, José, Khalifa, Kareem, and Leuridan, Bert (2013).

Pragmatism therefore offers a straightforward solution to a puzzle of this form, in which an explanation appears to be both bad and good. However, part of what makes the dormitive virtue case interesting is that some of its bad features appear to preclude it from *ever* explaining, not just failing to explain in a particular context. For example, even committed pragmatists struggle to accommodate absolute circularity in explanation, and one of the primary concerns about the dormitive virtue is that it may be circular.[24] So, although embracing pragmatism will go some way toward resolving the dormitive virtue puzzle, it will not do so perfectly. Furthermore, an interesting aspect of this puzzle is that positive verdicts on DV General explanation come not only from pragmatist philosophy of science, but also from literatures that are not motivated by pragmatism and that take connections between explanation and metaphysics seriously, including work on grounding and the metaphysics of dispositions. As we will see, one important task in resolving the dormitive virtue puzzle is to clarify the metaphysics at work in this case. Accordingly, I am interested in exploring responses to this puzzle that are not purely funded by pragmatism about explanation, although, as we will see, one of the benefits of the approach I will eventually recommend is that it can accommodate some central pragmatist insights.

Let us now turn to contemporary views that appear to permit DV General explanations. One source is dispositional essentialism, the position that (at least) fundamental properties are essentially dispositional.[25] Dispositional essentialists hold that these fundamental dispositions are the basis of natural modality and explain the laws of nature.[26] On this approach the explanatory work we may traditionally expect to be performed by the laws of nature is performed instead by these essentially dispositional properties, which provide a metaphysical and explanatory foundation for modality. The dispositions have no further categorical basis, as they are metaphysically and explanatorily fundamental. As Bird puts it in his preferred language of "potencies," "the existence of regularities in nature, the truth of counterfactuals, and the possibility of explanation are explained by potencies."[27]

Consider a concrete example. Bird discusses Reichenbach's famous comparison between the accidental fact that there is no ten-ton sphere of gold and the non-accidental fact that there is no ten-ton sphere of uranium. According to Bird, the latter is entailed and explained by uranium's dispositions to chain react

[24] For instance, even for Van Fraassen an answer to a why-question cannot be content of the question itself.

[25] For example see Bird, Alexander (2007) and Ellis, Brian (2001) The "at least" is in parentheses because dispositional essentialists differ over whether all properties are essentially dispositional (monism, Bird's position) or only certain properties.

[26] I say "natural modality," but for Bird the laws of nature are metaphysically necessary so he acknowledges no distinction between natural and metaphysical modality.

[27] Bird, Alexander (2007) pg 200.

and explode before it reaches that weight.[28] Here we can see an explanation of the structure of DV General:

> DV General: F reliably φs because F has a φ-ing virtue.
>
> Uranium: Uranium chain reacts and explodes before it reaches a ten-ton weight because it has a disposition to chain react and explode before it reaches that weight.

Some regard these dispositional explanations as causal explanations. However, others have framed the explanatory relationship between fundamental dispositions and the patterns in events that they explain in terms of grounding.[29] This leads to another contemporary source of DV General explanation. Proponents of grounding take grounding to be either a form of explanation, in its *unionist* variety, or a relationship between facts that supports, or *backs* explanation, in its *separatist* variety.[30] On either version wherever we have grounding, we have explanation, and that connection between grounding and explanation is one of its characteristic features. To stick with dispositional essentialism for the moment, most accounts of grounding can accommodate grounding between fundamental dispositions and patterns in events or laws of nature. Furthermore, because grounding explanations need not take us to the most fundamental explanatory basis for the explanandum in order to explain, we need not endorse dispositional essentialism to permit patterns in events to be grounded and hence explained by dispositions. Dispositions may have categorical bases and still feature in grounding explanations, and some grounding theorists explicitly discuss cases in which grounding occurs between facts about dispositions and facts about patterns in events. For example, Rosen discusses the idea that the fact that a ball is blue may be grounded in, and hence explained by, the ball's dispositions to reflect light in certain ways such that it appears blue.[31] Grounding explanations are notable for being extremely fine-grained such that the relata of a grounding explanation can be very close, as is evident in examples such as the grounding and hence explanation of the fact that the paint is red by the fact that the paint is scarlet, or the grounding and hence explanation of the fact that a person is a bachelor by the fact that they are an unmarried man. Accordingly, the apparently very close relationship between the dormitive virtue and opium's effects need be no barrier to a grounding explanation in this case.

[28] Bird, Alexander (2005) pg 357. Some have argued that permitting DV General explanation is a problem for dispositional essentialism. For example, Kimpton-Nye argues that a canonical version of dispositional essentialism permits DV General explanation, and defends an alternative version that does not, in Kimpton-Nye, Samuel (2021).

[29] For example, Jaag frames dispositional essentialism in terms of grounding in Jaag, Siegfried (2014). Note that the grounding literature was in a nascent stage at the time when many dispositional essentialists were developing their views.

[30] This way of describing the distinction comes from Raven, Michael (2015).

[31] Rosen, Gideon (2010) pg 126.

Furthermore, grounding explanations are non-causal, so the concern that virtues do not cause their manifestations does not apply here.[32]

Not all grounding theorists and not all of those who endorse a theory of dispositions endorse DV General explanations. However, many do, and in light of the worries raised in Section 2 these cases generate a puzzle. It seems obvious that DV and DV General explanations are problematic. But theories of grounding and dispositional essentialism provide good precedent for taking DV General explanations to be legitimate. Accordingly, the dormitive virtue case pulls us in two directions, and raises troubling questions. If we take the negative considerations against DV General explanations seriously, should we reject grounding and dispositional essentialism? Alternatively, if we take grounding and dispositional essentialism seriously, should we reject Molière's sarcastic verdict on this case?

This puzzle is not just an interesting puzzle about a single case, but asks us more generally to consider *how we think about explanation*. What role should intuitive responses to particular cases play when building a theory of explanation? Does it make sense to develop a theory of explanation and apply it top down to cases, regardless of the counterintuitive implications for some of those cases? What is the appropriate level of back and forth between a theory of explanation and explanatory practice? To what extent should we take pragmatics seriously? How should connections between explanation and metaphysics play into decisions about the viability or otherwise of explanations?

In what follows I will recommend an approach to the dormitive virtue case that is illuminating and offers sensible answers to at least some of these questions.

4 A view of explanation: CPN Backing

Contextualist Pluralist Non-Realist Backing (hereafter CPN Backing) is a backing model of explanation.[33] The central insight of backing models is that explanations are supported by underlying *backing relations* (or *backers*), and that we explain by reporting on these relations. On this approach explanation itself is a relation between propositions (or sentences), divided into two parts, an explanans and explanandum. For the explanation to succeed, the explanans must give information about something standing in a backing relation to whatever is described in the

[32] At least on a standard understanding of "causation." Wilson argues that grounding is a distinctively metaphysical form of causation in Wilson, Alastair (2018). On this approach DV General explanations may be legitimate, but they do not function as ordinary causal explanations.

[33] This section discusses a view of explanation developed by the author in Taylor, Elanor (2018), (2020), and (2023). For articulation and defence of backing models, see Audi, Paul (2015), Jaegwon Kim in a range of venues including Kim, Jaegwon (2005), Ruben, David-Hillel (1990), Jonathan Schaffer in a range of venues including Schaffer, Jonathan (2015), and Wilhelm, Isaac (2021)

explanandum. Take a simple causal explanation as an example, in which I explain a car crash by giving information about the brake failure that caused it:

The car crashed because the brakes failed.

We can divide this into two parts:

Explanandum: The car crashed.
Explanans: The brakes failed.

The explanandum gives information about the event that needs to be explained, and the explanans gives information about a cause of the event described in the explanandum. The explanation succeeds, at least in part, because the explanation is supported by and gives information about the backing relation of causation which obtains between the event described in the explanans and the event described in the explanandum.

Backing models vary on a range of different aspects, some of which will be significant for this discussion. These include the number, character, and unification of the backers, the extent to which the model accommodates contextual and pragmatic aspects of explanation, whether explanation reports on the relata of the backing relation or the relation itself, and the relationship between the structural and formal features of backers and the structural and formal features of explanation.

CPN Backing builds on this rough sketch of a backing model, but differs from extant backing models on a number of dimensions. I will sketch the central features of CPN Backing as a series of principles:

(1) Explanation is a relation between two (sets of) propositions, the explanans and explanandum.
(2) Explanations are backed by dependence relations that are not themselves explanations, but that can support explanation.
(3) The explanans of a successful explanation gives information about whatever that which is described in the explanandum depends on.
(4) There are many different backers including causation, grounding, mereological relations, conceptual relations, mathematical relations, logical relations, and motivational relations.
(5) There is some mirroring between the structural and formal features of backers and of explanation.
(6) Context determines which backers, and hence which explanations, are explanatorily appropriate. Features of context include the needs of the audience for the explanation, the activity at hand, and the background information available to those involved.
(7) Some backers, such as causation and grounding, are mind-independent, which means that the relation in general does not rely for its existence on human

> thought. Other backers, such as conceptual dependence, are mind-dependent, in that the relation in general does rely for its existence on human thought.

A primary, and controversial, difference between CPN Backing and more traditional backing models is the wider range of backers. Many backing models are pluralist in that they permit more than one backer, but typically these are restricted to causation and grounding. CPN Backing permits a variety of backers, including causation and grounding as well as conceptual relations, motivational relations, and mathematical relations. On traditional backing models the backers are highly unified, and for some this unification is reflected in the fact that the structure of backing is captured by the formalism of structural equation models.[34] CPN Backing does not posit a unified formalism for backing, and overall I endorse looser connections between explanation and backers than are posited by extant models. A further distinctive feature of CPN Backing pertaining to the unification or otherwise of backers is the non-realism. Backing models tend to be realist, in that backing relations are all mind-independent, worldly forms of metaphysical determination. This makes sense of the standard restriction to causation and grounding, as these are canonical forms of metaphysical determination. CPN Backing permits these worldly, metaphysical backers but also permits backers that are mind-dependent, including conceptual dependence, motivational dependence, logical relations on a conventionalist view of logic, and so on.[35] The non-realist aspect of CPN Backing is less significant for the dormitive virtue puzzle than some other features. But overall, this view permits a wider range of backers than is standard in backing models.

A further distinctive aspect of CPN Backing is the contextualism. Traditionally, backing theorists tend to endorse a robustly metaphysical approach to explanation, and leave aside issues about context and the pragmatics of explanation. This is not to say that backing theorists reject the idea that there *are* pragmatic and contextual aspects to explanation, but rather that they do not build these into their model of explanation. CPN Backing, on the other hand, places these issues at the heart of the view. A variety of different dependence relations may serve as backers, and context will determine which it is reasonable to cite in an explanation. For example, a metaphysics seminar room will be a more appropriate place to offer an explanation backed by grounding than almost any other context. Alternatively, the kind of information we desire from an explanation will be determined by factors such as whether we want to explain an event from an engineering perspective, or to forensically assign blame, and so on.

[34] As in Schaffer, Jonathan (2015), Wilhelm, Isaac (2021), and Wilson, Alastair (2018).

[35] For further discussion of this aspect, see Taylor, Elanor (2020).

I will leave a few important issues aside here. These include the formal features of backers. As stated above, on this model the formal features of backers reflect features of explanation and vice versa, which include irreflexivity, asymmetry, and hyperintensionality. However, unlike traditional backing theorists I take this mirroring between explanation and backing to obtain only at the level of instances that support specific explanations. On this approach, for instance, the combination of explanation's irreflexivity and causation backing explanations does not preclude the possibility of reflexive causation, so long as those instances of causation do not back explanations. I will also leave aside questions about the nature and extent of the unification of backers, beyond noting that I reject the view that the unification of backers is reflected in their subsumption under the formalism of structural equation models.

5 Taking on the dormitive virtue puzzle

The dormitive virtue puzzle is generated by competing positive and negative considerations about dormitive virtue explanation, and other explanations of the form DV General. On the negative side these explanations seem bad for a number of reasons, surveyed in Section 2. On the positive side, from a pragmatic perspective DV General explanations seem to give us at least some useful explanatory information, and, as surveyed in Section 3, well-motivated views of grounding and dispositions appear to permit explanations with this structure.

CPN Backing does not offer a definitive solution to this puzzle in that it does not provide a verdict on whether the dormitive virtue explanation is good or bad. This is because, as we will see, the dormitive virtue puzzle is generated by substantive questions about metaphysics and about explanatory context which an account of explanation alone cannot, and ought not, settle. But CPN Backing offers a useful diagnostic approach in that it clarifies what questions must be addressed in order to solve the puzzle, and accommodates and contextualizes a range of positive and negative responses to the case.

CPN Backing provides two criteria for explanation that are particularly salient to the dormitive virtue puzzle. The first is that an explanation must report on an instance of *dependence*—the backer. The second is that the dependence reported upon must be *contextually appropriate*. The first takes us to questions about the metaphysics operative in each case of explanation. The second takes us to considerations about the pragmatics of explanation.

Let us apply these criteria to the original case:

DV: opium reliably induces sleep because it has a dormitive virtue.

We must ask whether this explanation reports on an instance of dependence, and if so, whether that form of dependence is contextually salient. The first question cannot be settled by an account of explanation alone, because it is a substantive metaphysical issue. For example, if there is a power in which the pattern in opium's effects is grounded and DV reports on this grounding, then the first criterion is met. If there is no such power, and the dormitive virtue simply is the pattern in opium's effects, then the appropriate dependence does not obtain and the attempt at explanation fails. However, an instance of grounding is not the only way in which the dependence criterion can be met. An instance of causal dependence, such that the pattern is caused by the dormitive virtue, or an instance of conceptual dependence, such that there is a conceptual relationship like analysis or explication between the concept "dormitive virtue" and "reliably inducing sleep," can also meet the dependence requirement.[36] As before, whether these dependence relations obtain is a matter of background metaphysics.[37] If there is no dependence, then there is no explanation in this case.

The second criterion becomes relevant once we have established that at least one dependence relation is cited by the explanation. Then we must turn to the context of the request for explanation and ask whether the dependence cited is appropriate to that context. This contextual aspect provides insight into how, even if the dependence criterion is met, the dormitive virtue puzzle may still arise. Putting issues about the pantomime aspect of the scene in Molière's play aside, DV arises in a *clinical* context. Typically, in clinical contexts the explanations we seek of regularities are causal explanations that give information about the mechanism through which the effect is reliably obtained. McKitrick reflects on this fact when she says, "More ought to be said about why opium causes sleep, and in fact, we can say more: opium contains alkaloids such as morphine which, being structurally similar to the body's naturally occurring peptides, bind to opiate receptors in the brain, causing sleep."[38] Causal-mechanical explanations are not the only explanations appropriate to clinical contexts, but they are paradigm clinical explanations, not least because they facilitate causal-mechanical interventions. Accordingly, even if DV meets the dependence criterion, it may for good reason not meet the criterion of contextual appropriateness.

However, part of the puzzle was dealing with *contrasting* intuitions about this case. CPN Backing tells us that if we are looking for a clinical explanation, DV might meet

[36] As mentioned in the exposition of CPN Backing in Section 4, permitting backers of this non-realist, conceptual sort is a more controversial aspect of CPN Backing, and it would be rejected by more traditional realist backing theorists. See discussion in Taylor, Elanor (2020). However, those who endorse CPN Backing need not endorse the view that there are such conceptual explanations, or that DV is one.

[37] For ease of expression I have characterized the work of identifying dependence relations as "metaphysics," but this is not perfectly accurate given that CPN Backing permits non-realist backing.

[38] McKitrick, Jennifer (2003) pg 349.

the dependence criterion while still failing for contextual reasons. But then what do we do with the competing judgment that DV is ok after all, and that commitment to grounding and dispositional essentialism commits us to DV General explanation? Here the contextual aspect of CPN Backing can again play a useful role. The positive verdicts on DV and DV General explanations came from metaphysics, a distinctive explanatory context in which explanations in terms of ground and power are regarded not only as legitimate, but also as deeper and more complete than practical alternatives such as causal-mechanical explanations.[39] Even if DV does report on dependence, we can accommodate both negative and positive responses to the case by acknowledging that the dependence it reports on may not be explanatorily appropriate in its original clinical context (a source of negative verdicts), even if it is appropriate in the metaphysics seminar room (a source of positive verdicts).

Paying attention to these contextual factors also provides some insight into the comedic aspect of Molière's scene. There is a long history in comedy of getting laugh moments out of strictly adequate but contextually inappropriate explanations. For example, consider this explanation given in an episode of 1980's sitcom *Police Squad*:

(Detective Frank Drebin and forensic scientist Ted Olsen examine a rock that was thrown through a window.)

DREBIN: Where'd it come from?
OLSEN: It's very interesting. I have a theory about that. As you know, Frank, billions of years ago our Earth was a molten mass. But for some reason, not understood by scientists, the Earth cooled, forming a crust, a hard igneous shell. What we scientists call "rock."[40]

This is an excellent explanation in a way but pragmatically disastrous, hence its comedic impact. And something similar may be true of DV, even if it meets the dependence criterion.

Let us now consider how to address the dormitive virtue puzzle from a particular perspective. Say that I am a dispositional essentialist and I permit facts about patterns in events to be grounded in and hence explained by facts about dispositions, but I also have reservations about Molière's original case. How should I reconcile these competing views? I have a range of options. The first is to claim that Molière and others were wrong about DV, and that it is not a bad explanation after all. On this approach I reject the long-standing negative verdict as misguided,

[39] Consider Fine's claim that grounding is the "ultimate form of explanation" in Fine, Kit (2001) pg 16, or the dispositional essentialist idea that dispositions are metaphysically and explanatorily fundamental.

[40] *Police Squad* Season 1 Episode 5, IMDB: https://www.imdb.com/title/tt0676271/.

rather than trying to understand why so many people endorsed it and laughed along with Molière. Another option is to argue that despite apparent similarities the structure of Molière's explanation does not mirror the structure of the explanation countenanced by the dispositional essentialist. For example, I could follow some early modern commentators in arguing that only a few fundamental powers can play this explanatory role, and so that nothing at the coarse-grained level of the dormitive virtue could do this work. However, I am then left with further worries about explanations with DV General structure that do cite fundamental powers, which are permitted by my view. A third, and better, option is to make use of the contextualist and pluralist resources of CPN Backing. I may, as described above, judge that Molière's explanation was a fine explanation metaphysically speaking because it reports on a dependence relation, but that it was not appropriate for a clinical context. CPN Backing provides resources to justify laughing at Molière's medical student for giving the wrong kind of explanation rather than for failing to give an explanation at all, which seems like a good option for the dispositional essentialist who still wants to enjoy the fun.

Let us now return to the concerns identified in Section 2 about distance, mystery, and causation. The worry about distance is straightforwardly addressed by the dependence criterion for explanation. If DV reports on a dependence relation, then according to CPN Backing there is enough distance for explanation. The other concerns, about mystery and causation, are not so straightforwardly addressed by CPN Backing, which is appropriate because they are generated by substantive questions about metaphysics. CPN Backing tells us about how these metaphysical issues factor into the legitimacy of the explanation, but it does not resolve the metaphysical issues themselves. Consider the worry that the dormitive virtue is mysterious. This points to broad questions about scientific realism in that it requires us to consider when, if ever, it makes sense to posit unobservable entities in order to explain patterns in events. The third set of concerns about causation will also depend on the background metaphysics, so CPN Backing does not offer a straightforward solution but does issue some desiderata for an explanation. The explanation must cite a dependence relation, so if causal exclusion precludes this then the attempt at explanation will fail. Furthermore, on CPN Backing it does not follow from the fact that some causes can back explanations that all causes do, so this approach can make space for views such as the idea discussed by Sober and Shapiro that there may be a cause in this case without an explanation.[41] Overall, CPN Backing tells us what is required from the metaphysics for the explanation to succeed, without inappropriately generating verdicts on the metaphysics.

[41] Shapiro, Larry, and Sober, Elliott (2007) pg 19.

6 Reflections and implications

The dormitive virtue puzzle is that there are good, well-motivated reasons for thinking that dormitive virtue explanation is good, and that there are good, well-motivated reasons for thinking that that dormitive virtue explanation is bad. I have discussed some considerations in favor of each side and shown that a particular approach to explanation, CPN Backing, can help us to steer through this puzzle. CPN Backing offers a straightforward set of criteria for resolving this puzzle, and ways to accommodate lingering positive and negative intuitions about the case.

Applying CPN Backing to the case of the dormitive virtue reveals a variety of different responses to the puzzle generated by this case. However, CPN Backing does not tell us which view to adopt. Does this mean that CPN Backing has failed to resolve the puzzle? In short, no. Of course, one can take any account of explanation and impose it top-down onto the dormitive virtue case. For instance, I could adopt a strictly causal view of explanation and resolve the puzzle by deciding whether the dormitive virtue explanation meets the criteria given in my account, ignoring further considerations about the intuitive pull of one consideration over another. However, although this is a pleasingly simple way to proceed and generates a straightforward verdict, this kind of approach risks over-simplifying this complex, historic case. The dormitive virtue explanation is rich in metaphysical and contextual detail, and requires an approach that takes both seriously. CPN Backing displays the attention to context characteristic of pragmatism about explanation, but without the anti-realism also characteristic of pragmatism about explanation, which denies robust, systematic connections between explanation and metaphysics. By combining the idea that explanation is often importantly tied to metaphysics with attention to the contextual aspect of explanation, CPN Backing offers resources to help us to take metaphysics and pragmatics equally seriously when forming judgments about this case. In doing so, CPN Backing offers a range of resources for not only arriving at a verdict about the case, but also for dealing with residual responses to and intuitions about dormitive virtue explanation.

We began with the puzzle that dormitive virtue explanation seems laughably bad, but also appears to be permitted by contemporary theories of grounding and dispositions. Resolving this puzzle required us to examine connections between explanation and metaphysics, and explanatory context. CPN Backing is an approach to explanation that takes both of these aspects seriously, and shows that the dormitive virtue case is not so puzzling after all.[42]

[42] With thanks to Michael Della Rocca, Katie Robertson, Alastair Wilson, and two anonymous referees for helpful comments. Particular thanks to my mother, Valerie Kaye, for her patience as I wrote this chapter in her apartment during a lengthy pandemic quarantine.

References

Achinstein, Peter (1983) *The Nature of Explanation*. Oxford University Press.

Audi, Paul (2015) “Explanation and Explication.” In Daly, Chris ed., *The Palgrave Handbook on Philosophical Methods*, 208–230. Palgrave Macmillan.

Bird, Alexander (2005) “The Dispositionalist Conception of Laws.” *Foundations of Science* 10: 353–370.

Bird, Alexander (2007) *Nature’s Metaphysics*. Oxford University Press.

Díez, José, Khalifa, Kareem, & Leuridan, Bert (2013) “General Theories of Explanation: Buyer Beware.” *Synthese* 190(3): 379–396.

Ellis, Brian (2001) *Scientific Essentiallism*. Cambridge University Press.

Fine, Kit (2001) “The Question of Realism.” *Philosophers’ Imprint* 1: 1–30

French, Steven (2019) “Defending Eliminative Structuralism and a Whole Lot More.” *Studies in History and Philosophy of Science Part* A 74: 22–29.

Glanvill, Joseph (1665) *Scepsis Scientifica, or, Confest Ignorance, the Way to Science: In An Essay of the Vanity of Dogmatizing, And Confident Opinion*. E. Coates.

Goodman, Nelson (1983) *Fact, Fiction and Forecast* (4th edition). Harvard University Press.

Hutchison, Keith (1991) “Dormitive Virtues, Scholastic Qualities, and the New Philosophies.” *History of Science* 29(3): 245–278.

Jaag, Siegfried (2014) “Dispositional Essentialism and the Grounding of Natural Modality.” *Philosophers’ Imprint* 14 (34): 1–21

Kim, Jaegwon (2005) *Physicalism, or Something Near Enough*. Princeton University Press.

Kimpton-Nye, Samuel (2021) “Reconsidering the Dispositional Essentialist Canon.” *Philosophical Studies* 178: 3421–3441.

Lange, Marc (2017). *Because Without Cause*. Oxford University Press.

Lewis, David (1986) “Causal Explanation.” In *Philosophical Papers Volume II*, 214–240. Oxford University Press.

Locke, John (1689) *Essay Concerning Human Understanding*. Clarendon Press.

Malebranche, Nicolas, English translation by Olscamp, Paul J. (1997) *The Search After Truth*. Cambridge University Press.

McKitrick, Jennifer (2003) “The Bare Metaphysical Possibility of Bare Dispositions.” *Philosophy and Phenomenological Research* 66(2): 349–369.

Molière, Jean Baptiste, English translation by Waller, A. R. (1926) *The Plays of Molière in French*. J. Grant. (8 Volumes).

Ott, Walter (2009) *Causation and Laws of Nature in Early Modern Philosophy*. Oxford University Press.

Prior, Elizabeth W., Pargetter, Robert, & Jackson, Frank (1982) “Three Theses about Dispositions.” *American Philosophical Quarterly* 19(3): 251–257.

Raven, Michael (2015) “Ground.” *Philosophy Compass* 10(5): 322–333.

Rosen, Gideon (2010) “Metaphysical Dependence: Grounding and Reduction.” In Hale, Bob, & Hoffmann, Aviv (eds.), *Modality: Metaphysics, Logic, and Epistemology*, 109–135. Oxford University Press.

Ruben, David-Hillel (1990) *Explaining Explanation*. Routledge.

Schaffer, Jonathan (2015) “Grounding in the Image of Causation.” *Philosophical Studies* 173: 49–100

Shapiro, Larry, and Sober, Elliott (2007) “Epiphenomenalism: The Dos and the Don’ts.” Machamer, Peter, & Wolters, Gereon (eds.), *Thinking about Causes*, 235–264. Pittsburgh: University of Pittsburgh Press.

Skow, Bradford (2014) “Are There Non-Causal Explanations (of Particular Events)?” *British Journal for the Philosophy of Science* 65(3): 445–467.

Sober, Elliott (1984) *The Nature of Selection: Evolutionary Theory in Philosophical Focus*. MIT Press.

Strevens, Michael (2008) *Depth: An Account of Scientific Explanation*. Harvard University Press.

Tahko, Tuomas, E. (ed.) (2011) *Contemporary Aristotelian Metaphysics*. Cambridge University Press.

Taylor, Elanor (2018) "Against Explanatory Realism." *Philosophical Studies* 175(1): 197–219.

Taylor, Elanor (2020) "Backing Without Realism." *Erkenntnis* 87: 1295–1315.

Taylor, Elanor (2023) "Explanatory Distance." *British Journal for the Philosophy of Science* 74(1): 221–239.

Thomasson, Amie (2014) *Ontology Made Easy*. Oxford University Press.

Van Fraassen, Bas (1980) *The Scientific Image*. Oxford University Press.

Vetter, Barbara (2015) *Potentiality*. Oxford University Press.

Wilhelm, Isaac (2021) "Explanatory Priority Monism." *Philosophical Studies* 178: 1339–1359.

Wilson, Alastair (2018) "Metaphysical Causation." *Noûs* 52(4): 723–751

Woodward, James (2003) *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

15

The Explanatory Role Argument and the Metaphysics of Deterministic Chance

Nina Emery

1 Introduction

It is becoming increasingly common for metaphysicians and philosophers of science to claim that there are non-trivial *objective probabilities*—or *chances*—in worlds where the fundamental laws are deterministic. Call such probabilities *deterministic chances*.[1]

Some arguments for deterministic chance begin by collecting platitudes about the chance-concept and then asking whether those platitudes—or a sufficient subset of them—apply to entities that exist in worlds where the fundamental laws are deterministic (Glynn 2010). Other arguments for deterministic chance begin by developing a full-blown metaphysical theory of chance and then seeing whether chances, according to that theory, arise in worlds where the fundamental laws are deterministic (Loewer 2001). This paper takes up a different sort of argument—one that purports to establish that chances play an important role in our best scientific theories *even if* the fundamental laws are deterministic. Whether we have good reason to posit the existence of things that *don't* play an important role in our best scientific theories may be an open question. But surely we ought to accept the existence of those things that *do* play such a role. So if chances play an important role in our best scientific theories even when the fundamental laws are deterministic, then we ought to accept the existence of deterministic chances.

More specifically, this paper is concerned with the *explanatory role argument for deterministic chance*. According to this argument, certain probabilities play an important explanatory role in our best scientific theories even if the fundamental laws are deterministic, and in order for those probabilities to play the relevant explanatory role, they must be objective probabilities. Therefore, the argument concludes, there are such things as deterministic chances.

The explanatory role argument was central to David Albert's influential discussion of classical statistical mechanics in *Time and Chance* (Albert 2000), and has

[1] I will use the terms "objective probability" and "chance" interchangeably throughout.

Nina Emery, *The Explanatory Role Argument and the Metaphysics of Deterministic Chance* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Nina Emery 2024.
DOI: 10.1093/oso/9780192862945.003.0016

reappeared regularly in the literature since then (Meacham 2005, Maudlin 2007, Emery 2015). Consider, for instance, the following quote from Barry Loewer:

> Physicists insist on producing explanations and laws involving probabilities even though the systems to which they apply are considered to be deterministic. The problem is that it is very difficult to see how these probabilities can play the roles that they are called upon to play in reliable predictions, explanations, and laws if they are merely subjective. (Loewer 2001, 610)

At the same time, the explanatory role argument is rarely, if ever, spelled out in detail and, perhaps because of this, its consequences have not been fully appreciated. In this paper I first present the explanatory role argument and then show that insofar as the argument is successful it places at least two important constraints on our further metaphysics of deterministic chance. As the discussion will demonstrate, the important explanatory role that deterministic chances are supposed to play involves higher-level macrophysical patterns, which are taken to require a probabilistic explanation, even though the underlying microphysical states may well be explained in an entirely deterministic way. In this way, a careful examination of the explanatory role argument illustrates an important application of distinct explanatory levels, as well as some of the potential consequences and constraints involved in providing probabilistic explanations of higher-level phenomena when the underlying physics is deterministic.

Before we begin, it is worth recognizing that, at least on the face of it, non-trivial deterministic chances are surprising. If the fundamental laws are deterministic then the laws together with the complete state of the world at any time fully determine what will happen at every other time. This gives rise to an initial challenge for any account of deterministic chance: how can there be some objective probability regarding what will happen, if what will happen is fully determined?

Although this initial challenge is philosophically interesting in its own right, I won't discuss it in any detail here. Instead I will simply assume in what follows that this sort of worry, however serious, is not so serious as to make deterministic chance a non-starter.[2] My goal here is not to convince those who don't believe in deterministic chance to change their minds. My goal, rather, is to identify some consequences that follow insofar as one thinks that there are deterministic chances, and specifically insofar as one thinks there are deterministic chances *because of the explanatory role argument*. Philosophers who endorse that argument, I claim, should tread carefully when it comes to accepting any of the existing

[2] Although this assumption contradicts Lewis 1986a and Schaffer 2007, it finds a great deal of support in the recent literature. For specific discussion of the challenge, see Glynn 2010, Handfield and Wilson 2014, and Emery 2015. For further work on deterministic chance in general, see Ismael 2009, Briggs 2010, and Demarest 2016.

metaphysical analyses of deterministic chance, for those analyses may not do the work for which they were designed.

2 The explanatory role argument for deterministic chance

You're hurrying around your apartment one afternoon getting ready to leave for vacation. You lock the windows and take out the trash, but—oops!—you leave a nice, ripe banana sitting in the middle of your kitchen counter. When you come back two weeks later, the banana has rotted.

According to classical mechanics, things could have been different. If all of the particles in the banana, the kitchen counter beneath it, and the air surrounding it had been arranged just right when you closed the door on your way out, you might have come home to a still perfectly ripe banana. Or even a banana that had become quite green.[3]

This fact about bananas is part of a more general fact about the temporally asymmetric patterns that we observe in the macrophysical world around us. According to classical mechanics, all of these temporally asymmetric patterns can in fact happen in reverse. Bananas can, contra ordinary experience, get increasingly green and firm. Broken vases can pull themselves back together and hop back up onto the edge of nearby tables. Ice cubes in a glass of lukewarm water can get bigger while the water around them gets increasingly warm. All of these strange behaviors have something in common: they involve a certain physical property—for our purposes we can just refer to it as the *entropy*—of the system in question decreasing over time. Scientists call these sorts of behaviors *anti-entropic*. Usually, the systems that we observe (specifically the closed systems that we observe that are not already at equilibrium) increase in entropy over time. But according to classical mechanics, anti-entropic behavior is possible.

The possibility of anti-entropic behavior raises an immediate explanatory burden: given that anti-entropic behavior is possible, why don't we observe it happening? And the standard way of meeting this explanatory burden is to appeal to probability.[4] Why don't we see bananas grow greener or vases un-break or ice

[3] Here I am following in the tradition of Albert (2000) in assuming that classical statistical mechanics applies universally. This tradition involves extending the theory significantly beyond the sorts of systems that physicists are actually able to straightforwardly model using classical statistical mechanics. Those who are skeptical of this extension are welcome to use a different example. Imagine, for instance, that I left for vacation with the door open between my bedroom (which was fairly cold) and my kitchen (which was fairly warm) and came home to find the whole space at roughly the same temperature. According to classical statistical mechanics it is possible for me to have returned to find the bedroom even colder and the kitchen even warmer than it had been when I left.

[4] So, for instance, Loewer writes, "The usual statistical mechanical explanation of why ... we never encounter a block of ice that behaves [in an anti-entropic way] is that the probability of a microstate that evolves in the first way is vastly greater than the probability of a microstate that behaves in the second" (Loewer 2001, 611). As examples of other philosophers who claim that this is the standard explanation see Strevens (2000, section 2), and Meacham (2010).

cubes get bigger? Because although all of these processes are physically possible, they are extremely unlikely. It is the introduction of these probabilities that marks the shift from classical mechanics to classical statistical mechanics.

And notice that insofar as we take this standard way of explaining the absence of anti-entropic behavior seriously, we are committed to the existence of deterministic chances. After all, the microphysical laws are still entirely deterministic.[5] Given the complete microphysical state of a closed system at a time, these laws will tell us the microphysical state at all other times. And whatever else we think about the probabilities that explain why bananas rot and ice cubes melt and broken vases remain where they are on the floor, those probabilities must be objective features of the world. No *subjective* feature of the world—no fact about what we happen to believe or have evidence for, or about the best way for creatures like us to reason—explains why bananas and ice cubes and vases behave the way that they do. Subjective features of the world might well play a role in explaining *why we expect* bananas and ice cubes and vases to behave a certain way. But they don't explain why that behavior *actually* occurs.[6]

Here, then, is one version of the explanatory role argument for deterministic chance:

P1 Probabilities explain statistical mechanical phenomena.[7]
P2 In order to explain statistical mechanical phenomena, the relevant probabilities must be objective.
C There are deterministic chances.

Other versions of the argument can be generated by considering other theories where the fundamental laws are deterministic but probabilities play a role in explaining patterns in the phenomena (e.g., Bohmian mechanics).

It is worth emphasizing that in order for P2 to be plausible, advocates of the explanatory role argument must be making use of a fairly specific notion of explanation. An explanation, for the advocate of this argument, cannot, for instance, just

A related claim is that only the explanatory power of deterministic chance can adequately capture the historical facts regarding the adoption of statistical mechanics instead of alternative theories in the second half of the nineteenth century. For more, see Strevens (2000).

[5] More carefully: if ours was a world in which classical mechanics held, then the microphysical laws would still be deterministic. In our actual quantum world, it is an open question whether the fundamental laws are deterministic or not. See the discussion at the end of the section.

[6] So, for instance, Wallace (2011) writes, "The problem with taking SM probabilities to be credences is that the probability distribution in statistical mechanics grounds objective features of the world" (203).

[7] According to a recent paper by Hicks and Wilson (2021), it is not the probabilities themselves that explain in these cases. Instead, the set-up of each specific situation in which statistical mechanics phenomena arise explains the outcome, and statistical mechanical probabilities "mediate" these explanations. I take it that what I say here and below can be adjusted to accommodate this alternative view.

be a linguistic act that allows us to understand some phenomena. If it were, then the probabilities that play a role in explaining statistical mechanical phenomena might well be subjective. Instead, the advocate of the explanatory role argument must be committed to what I will call a *metaphysically robust* notion of explanation. I won't attempt to give a full characterization of metaphysically robust explanations here. It will suffice for our purposes to note the following two points. First, a metaphysically robust explanation identifies *the reason why* the explanandum occurred. And second, paradigm examples of metaphysically robust explanations are explanations that identify the cause or the ground of the explanandum.[8]

As long as we understand explanation as it plays a role in the explanatory role argument as metaphysically robust explanation, P2 is unassailable. No subjective feature of the world can be the reason why bananas rot or ice cubes melt.[9]

But note that this way of understanding explanation has consequences for how we understand P1. As we saw above, the reason that advocates of the explanatory role argument give for endorsing P1 is that scientists use (1) to explain both B-S and B-P.

(1) The probability of any particular banana that has been left on the counter for two weeks rotting is very high.
B-S The banana that you left on the counter for two weeks rotted.
B-P Most bananas that are left on the counter for two weeks rot.

And more generally, scientists use the high probability of entropy increasing over time for a closed system not in equilibrium to explain the fact that most such systems behave in accordance with the second law of thermodynamics.[10] Given the notion of explanation required in order to make P2 plausible, advocates of the explanatory role argument must think that when scientists give probabilistic explanations of statistical mechanical phenomena, they are successfully identifying the *reason why* those phenomena occur. Scientists cannot merely be making use of the probabilistic explanation because it is particularly simple to use, or because it is

[8] Insofar as one does not endorse grounding per se but rather a range of more specific dependence relations (as in Wilson 2014), metaphysically robust explanation will track those relations. Metaphysically robust explanations can also track primitive governance relations, as discussed in Emery (2023).

[9] Note that subjective features of the world can of course be a part of the reason why *we believe* bananas rot and ice cubes melt or even the reason why *we predict* this behavior. The point here is that anyone who endorses the explanatory role argument must think that there is a type of explanation that goes beyond merely identifying the reason why we have certain beliefs or expectations. There will of course be some who refuse to accept this metaphysically robust sort of explanation, but they should not endorse the explanatory role argument for deterministic chance and thus are not the target of the discussion here.

[10] For a particularly clear defense of this point, see Meacham (2010), where he says that to deny this would be revisionary with respect to: (i) textbooks, (ii) people's experience of learning statistical mechanics, and (iii) the history of how statistical mechanics was accepted. He goes on to say: "And keep in mind how pervasive these revisions will need to be. many explanations in other fields, especially upper-level sciences, are often premised on thermodynamic phenomena."

easier to understand than alternative explanations. They must be putting forward the relevant probabilities as the reason why anti-entropic behavior does not occur.

This understanding of explanation makes P1 more controversial than it may at first have appeared. But remember that my goal here is not to convince anyone to accept the explanatory role argument. Rather it is to show that the explanatory role argument, if accepted, constrains one's metaphysics of deterministic chance in significant ways. I will turn momentarily to discussing the first of these constraints.

Before I do so, however, it is worth saying something about one initial worry that may arise about the argument. The initial worry is this: classical statistical mechanics is not one of our best scientific theories. We have decisive empirical evidence that shows that the fundamental laws of nature are not the laws of classical mechanics. Doesn't that make the argument above either invalid or a non-sequitur? It doesn't follow from the explanatory role argument that there are actually any deterministic chances. It only follows that there are deterministic chances in worlds in which classical statistical mechanics holds. But since we know that our world is not such a world, who cares?

In fact there are several reason to care, but I'll focus here on just this one: although classical statistical mechanics is not one of our best scientific theories, some quantum version of statistical mechanics is, and while it isn't certain that quantum statistical mechanics will involve deterministic fundamental laws, it is very much an open question. So the actual world may or may not turn out to have deterministic chances as supported by the explanatory role argument. Presumably then, we should try to understand what such quantum deterministic chances would be like, since they may in fact actually exist. As the presentation is less complex and less nuanced when thinking about the fundamental laws as the laws of classical mechanics, however, I will continue to do so here and throughout.[11]

3 The explanatory constraint

The first constraint that follows from the explanatory role argument should now be obvious. Insofar as you endorse the explanatory role argument for deterministic chance, you must think that deterministic chances explain statistical mechanical phenomena. More specifically, given the discussion above about the notion

[11] Here are two other reasons to care. First, it may be that statistical mechanical phenomena is screened off from the fundamental level as described in Wallace (2011). Then one should give the same account of how statistical mechanical probabilities work regardless of whether we are in a quantum or a classical world. Second, the role that probabilities play in Bohmian mechanics—a deterministic interpretation of quantum mechanical phenomena—is similar to the role that they play in classical statistical mechanics. Therefore, a better understanding of the explanatory role of probabilities in classical statistical mechanics may help shed light on our understanding of the explanatory role of probabilities in Bohmian mechanics.

of explanation required to make P2 plausible, advocates of the explanatory role argument are committed to thinking that:

> *The explanatory constraint.* Deterministic chances provide metaphysically robust explanations of statistical mechanical phenomena.

Or, to focus on a particular example, advocates of the explanatory role argument are committed to thinking that (1) is the reason why both (B-P) and (B-S) occur.

Although this constraint should be relatively obvious, it is also enough to create serious difficulties for certain metaphysical accounts of chance that have been proposed as metaphysical accounts of deterministic chance. In particular, it creates difficulties for Humean accounts of deterministic chance, including the best systems analysis put forward in Loewer 2001.

3.1 The explanatory constraint and actual frequentism

In order to see why this is the case, it helps to start with a particularly simple Humean account, the *actual frequency account.* According to the actual frequency account, the chance of an event of type E happening in a situation of type S is just the relative frequency with which events of type E actually occur in situations of type S. According to this sort of account, to say that the probability of any particular banana rotting is very high is just to say that the vast majority of bananas rot.

Actual frequency accounts are not often defended in the literature on chance, but they are worth discussing here for a few reasons. First, actual frequentism appears to be a key component of one of the more influential recent accounts of deterministic chance: Hoefer's Humean Objective Chance. According to Hoefer, "Chances are constituted by the existence of patterns in the mosaic of events in the world" (2007, 580) and some chances are "simply there, to be discerned, in the patterns of events" (564). This suggests that although not all chances are actual relative frequencies, some of them are. And insofar as they are, they cannot satisfy the explanatory constraint. Perhaps Hoefer can argue that the classical statistical mechanical chances are not just actual frequencies and that the chances that are just actual frequencies do not play any important explanatory role. But that will take some work.

Second, although it is of course difficult to give concrete evidence for such assertions, my sense is that insofar as there is a standard metaphysical account of objective probability among practicing scientists, it is actual frequentism.[12] The

[12] Just as an example, here is Richard Feynman from his *Lectures on Physics*: "By the 'probability' of a particular outcome of an observation we mean our estimate for the most likely fraction of a number of repeated observations that will yield that particular outcome" (Feynman et. al. 1963, 6–1).

argument given below, then, reveals a real tension at the heart of scientific practice vis-à-vis probabilities. The metaphysical account that scientists often give of chances and one of the roles that they often think chances play don't appear to be compatible.

Third and finally, the issues with adopting actual frequentism (in addition to the explanatory role argument for deterministic chance) help illustrate the issues that arise for more popular Humean views.

That there will be issues along these lines for the actual frequentist is easy to see. Insofar as one understands deterministic chance as actual relative frequency *and* one endorses the explanatory role argument, one faces immediate problems. For now (1) is just equivalent to (B-P). So in order to satisfy the explanatory constraint, (B-P) would have to provide a metaphysically robust explanation of itself. The reason that most bananas that are left on kitchen counters for two weeks rot is that most bananas that are left on kitchen counters for two weeks rot. This is unacceptable. Nothing can provide a metaphysically robust explanation of itself.[13]

Here is another way to put the point: insofar as one adopts the actual frequency account of deterministic chance alongside the explanatory role argument for deterministic chance, one is committed to at least one instance of Explanatory Symmetry:

> *Explanatory Symmetry*. There is some A such that A is the reason why A occurs.

But Explanatory Symmetry is unacceptable—or at least, it is such a significant cost that one presumably ought to revise one's other commitments if at all possible in order to avoid it. Call this the *explanatory symmetry challenge* for actual frequentism about deterministic chance.

Now, there are some moves that are available to the advocate of the actual frequency account in response to the explanatory symmetry challenge. For instance, one might try to claim that deterministic chances explain individual events but not patterns of events—so (1) explains B-S, but not B-P. But this would be problematic in the present context. For scientists *do* seem to use (1) to explain B-P. And insofar as one is willing to reinterpret what scientists are doing when they seem to be using (1) to explain B-P, then why accept P1 in the explanatory role argument to begin with?

Alternatively, the actual frequentist might try to avoid the explanatory symmetry challenge by saying that (1) explains B-P only indirectly. For instance, they might say that (1) explains B-S and all other individual events involving bananas

[13] Hajek puts forward a similar critique of the actual frequency account as an account of indeterministic chance. "Why do we believe in chances? Because we observe that various relative frequencies of events are stable; and that is exactly what we would expect if there are underlying chances with similar values. We posit chance in order to explain the stability of these relative frequencies. But there is no explaining to be done if chance just is actual relative frequency" (1996, 79).

rotting, and all those individual events, taken together, explain B-P. But although this avoids a commitment to Explanatory Symmetry, it still involves a commitment to something that is not much better. Since the actual frequentist thinks that (1) is equivalent to B-P, to take the route just described is to say that B-P provides a metaphysically robust explanation of the conjunction of all individual facts involving bananas rotting, and that conjunction in turn provides a metaphysically robust explanation of B-P. So, insofar as one defends an actual frequentist account of deterministic chance via this route, one can avoid commitment to Explanatory Symmetry, but will still be committed to at least one instance of Explanation Circularity:

> *Explanatory Circularity*. There is some A and B such that A is the reason why B and B the reason why A.

But again, Explanatory Circularity is either unacceptable, or such a significant cost that one ought to revise one's other commitments if at all possible in order to avoid it.

3.2 The explanatory constraint and Humean theories in general

All that by way of showing how the explanatory constraint is problematic for those who want to understand deterministic chance as actual relative frequency. Turn now to Humean accounts more generally. What these accounts have in common is the view that the chances are in some important sense "nothing over and above" the Humean mosaic—the actual distribution of non-modal entities throughout spacetime. David Lewis understood this as a claim about supervenience; the chances, on his view, supervened on the mosaic.[14] Recently, many Humeans have put the central claim in terms of grounding; what it is to be a Humean, on this view, is to think that the chances are grounded in the mosaic.[15] I will initially present the worry for Humeanism in terms of this grounding claim. But those who are skeptical of understanding Humeanism in this way should note that there is a version of the worry (which I will also present below) that targets even those who insist that Humeanism should be understood merely as a claim about supervenience.

So for now, start with the assumption that insofar as one adopts a Humean account of deterministic chance, one is committed to the view that deterministic chances are grounded in the Humean mosaic. More specifically, one is committed

[14] See Lewis (1986b, ix–x).

[15] See, for instance, Beebee (2000), Schaffer (2008), Loewer (2012), and the discussion in Emery (2020). These philosophers are often talking about Humeanism with respect to laws, but almost universally Humeanism about laws and Humeanism about chance go hand in hand.

to thinking that (1) is grounded in the Humean mosaic. And remember that explanations that identify the grounds of the explanandum are paradigm examples of metaphysically robust explanations. So insofar as one adopts a Humean account of deterministic chance, one is committed to the view that the Humean mosaic provides a metaphysically robust explanation of, e.g., (1). The Humean mosaic is the reason why (1) is what it is.

But now suppose that one is not only a Humean about deterministic chance but a Humean about deterministic chance who endorses the explanatory role argument. Then, as follows from the explanatory constraint, one is also committed to the view that (1) provides a metaphysically robust explanation of B-S and B-P. And B-S and B-P are part of the Humean mosaic. So (1) partly explains the Humean mosaic (in a metaphysically robust sense). So the Humean about deterministic chance who endorses the explanatory role argument is committed to the Humean mosaic providing a metaphysically robust explanation of (1) and (1) partially providing a metaphysically robust explanation of the Humean mosaic. They are committed, in other words, to there being at least one instance of Partial Explanatory Circularity.

> *Partial Explanatory Circularity*. There is some A and B such that A is the reason why B and B is part of the reason why A.

Perhaps Partial Explanatory Circularity is not as bad as Explanatory Circularity full stop, or Explanatory Symmetry, but it is still surely a consequence to be avoided. Indeed, I submit that insofar as partial explanatory circularity follows from the combination of Humeanism about chance and the explanatory role argument, one ought to give up either Humeanism about chance or the explanatory role argument.

Readers familiar with the recent literature on laws of nature will note that a prominent recent attempt to avoid a similar sort of explanatory circularity objection to Humeanism about laws of nature would seem to be of some help here. Humeans about laws, on the face of it, also seem to be committed to Partial Explanatory Circularity. In virtue of being Humeans they are presumably committed to the Humean mosaic being the reason why the laws are what they are. But there are also good reasons for thinking that the laws at least partially explain the Humean mosaic.

In his paper "Two Accounts of Laws and Time," Loewer (2012) argues that although Humeans are committed to a certain kind of explanatory circularity, that sort of circularity isn't problematic because the type of explanation involved changes as one navigates the circle. In particular, Loewer says that while the Humean mosaic metaphysically explains the laws, the laws scientifically explain parts of the mosaic. And it is fine for A to metaphysically explain B while B scientifically explains A. The analogous view in the case of chances is that the Humean mosaic metaphysically explains (1), while (1) partly scientifically explains the mosaic. And again, according to Loewer at least, there is no reason why A cannot metaphysically explain B while B partly scientifically explains A.

There has been quite a bit of active debate about whether Loewer's move is legitimate.[16] But for our purposes, what is important is to notice the way in which the explanatory constraint puts pressure on anyone who was hoping to use this maneuver to defend Humeanism about deterministic chance. For notice that whatever else we might mean by "scientific explanation," the explanatory constraint tells us that this sort of explanation has to be metaphysically robust. If A scientifically explains B then A is the reason why B occurs. It follows that the Humean about deterministic chance who endorses the explanatory role argument cannot avoid accepting Partial Explanatory Circularity as stated above as a consequence of her view.

Let me return to the promise made earlier to say something about how the worry above applies if you reject the currently popular move of understanding Humeanism in terms of grounding. First note that Partial Explanatory Circularity does not mention grounding directly at all—it is a claim about a more general kind of explanatory relation, metaphysically robust relation, of which explanations that identify grounds are paradigm cases. Understanding Humeanism in terms of a claim about grounding was only relevant insofar as it gave us a quick reason for endorsing part of the circularity in Partial Explanatory Circularity: insofar as you take the mosaic to ground the chances, then you think that the mosaic is the reason why the chances are what they are. Consider, then, a ground-free Humean, who thinks that Humeanism about chance is characterized by supervenience. The key question for such a Humean in the context of the worries raised here is: what is the reason why the chances are what they are? If the ground-free Humean answers that the mosaic is the reason why the chances are what they are, then she faces just as much of a problem with Partial Explanatory Circularity as did her ground-friendly colleague. Moreover, it seems surprising and strange for the ground-free Humean to resist this answer. If the mosaic is not the reason why the chances are what they are, according to the Humean, what exactly is the reason? It seems difficult, at best, to see how someone could both claim to be a Humean *and* claim that there is no reason why the chances are what they are.

I will leave the discussion of the way in which the explanatory constraint influences one's metaphysics of deterministic chance here. I don't claim to have shown that the explanatory constraint definitively rules out Humeanism about deterministic chance. What I have shown is that that constraint creates serious problems for such accounts. Without quite a bit more work, it is not at all clear that one can both endorse the explanatory role argument and also maintain a Humean account

[16] See Lange (2013), Hicks and van Elswyck (2014), Miller (2015), and Emery (2018), for further discussion.

of deterministic chance. Given the prevalence of Humean accounts of deterministic chance in the literature, and the fact that some prominent defenders of the explanatory role argument are Humeans (Albert 2000 and Loewer 2001), this is a substantive result.

4 A further constraint

The first constraint that the explanatory role argument places on the metaphysics of deterministic chance—the explanatory constraint—is fairly obvious, if also somewhat under-appreciated in the literature. The second constraint is considerably less obvious. Indeed, I am not entirely sure what form this constraint takes. I am confident, however, that there must be *some* further constraint. And I think there is good initial reason to suspect that this constraint is substantive—and in particular that, like the explanatory constraint, it creates problems for Humean accounts of deterministic chance. Moreover, the metaphysics of deterministic chance that is most straightforwardly compatible with this constraint is one that has seen little attention in the literature.

4.1 Why there must be a second constraint

To see why there must be a second constraint that follows from the explanatory role argument, compare the explanatory role argument for deterministic chance with the nearby explanatory role argument for indeterministic chance:

P1* Probabilities explain (indeterministic) quantum mechanical phenomena.
P2* In order to explain (indeterministic) quantum mechanical phenomena, the relevant probabilities must be objective.
C* There are (indeterministic) chances.

Here there is the same initial argument for P1* as there was for P1: scientists apparently endorse this premise, so we should too. But there's also a second, much stronger argument: if chances don't explain the sort of quantum mechanical phenomena in question, then nothing does; and something *must* explain that phenomena. To think otherwise would be to violate a key norm of standard scientific practice. In particular, it would violate the norm that says we ought not leave robust patterns—like the sorts of robust patterns that constitute quantum mechanical phenomena—unexplained.[17]

[17] I discuss this norm in the context of the argument from P1* and P2* to C* in detail in Emery (2017).

Note that there is no immediate back-up argument of this sort for P1. At least at first glance, there *are* alternative, non-probabilistic explanations of statistical mechanical phenomena. In particular, there are alternative, non-probabilistic explanations for statistical mechanical phenomena that *just* cite the actual initial microphysical conditions of the system combined with the deterministic fundamental laws. One might claim, for instance, that B-S is explained by (2):

(2) When you left the banana on the counter for two weeks, the system started off in initial microstate m_1 and (in combination with the fundamental laws) m_1 led deterministically to rotting.

And B-P is explained by (3):

(3) In most instances in which someone leaves a banana on the counter for two weeks, the system starts off in initial microstate m_1 or m_2 or … or m_n and (in combination with the fundamental laws) all of those initial microstates lead deterministically to rotting.

It follows that insofar as you endorse P1, you must think that the probabilistic explanation for statistical mechanical phenomena is in some important sense superior to these alternative, non-probabilistic explanations.[18] (And insofar, then, as scientists are claiming that (1) explains B-P and B-S they are endorsing (perhaps implicitly) a certain view about what makes some explanations superior than others, and the conditions under which a higher-level explanation is warranted or necessary.)

And notice that whatever else you want to say about the criteria by which probabilistic explanations for statistical mechanical phenomena are superior to alternative, non-probabilistic explanations, those criteria must be wholly objective. There are, of course, many ways that one explanation—and specifically, in this case, one metaphysically robust explanation—might be better than another. One metaphysically robust explanation might be better in the sense that it is easier to use. Another might be better in the sense that it is closer to what I was hoping for. What we are looking for here, however, is a set of criteria that make one metaphysically robust explanation better than another in the sense that licenses something like inference to the best metaphysically robust explanation. We are looking, in other words, for a set of criteria that allow us to determine which explanation correctly identifies the reason why the explanandum occurred. And in general (and certainly for the

[18] Strictly speaking, this could happen in one of two ways. First, the alternative, non-probabilistic explanations might not be genuine candidates for explaining statistical mechanical phenomena *at all*. Second, the explanation by way of deterministic chance might be in some sense superior, so that inference to the best explanation licenses our endorsement of the probabilistic explanation. I will assume the latter throughout what follows.

type of explanandum in question here), the reason why something happened is independent of the types of creatures we are and the way that we are situated in the world.

So the question under consideration here is: what is the (wholly objective) criterion by which the probabilistic explanations of statistical mechanical phenomena are better explanations than any alternative, non-probabilistic explanations? What, for instance, is the (wholly objective) criterion by which (1) is a better explanation of B-S and B-P than any alternative, non-probabilistic explanation like that given by (2) and (3)?

This is a hard question to answer, and I won't aim to answer it in full. In part, my goal here is just to demonstrate that there must be some criterion of this form, and that therefore those who put forward the explanatory role argument for deterministic chance alongside a fully developed metaphysics of chance should proceed with caution. Until this second criterion is fully understood, they cannot be confident that their theory of deterministic chance itself will allow deterministic chances to do the work for which they were introduced.

But I do have some suspicions about the form that this second criterion might take. In the next two subsections I will walk through two candidates for this criterion. The first, which comes out of the literature on higher-level explanation, though initially plausible, does not, I think, end up working (though it is instructive to see why). The second, I suspect, might be made to work, and, interestingly, insofar as it is, it is likely to make further problems for Humean accounts of deterministic chance.

4.2 Modal robustness

Start from the observation that there are two notable differences between (1), on the one hand, and (2) and (3), on the other, as candidate explanations of B-S and B-P. First, (1) merely makes B-S and B-P likely, whereas (2) entails B-S, and (3) entails B-P. This is a consideration that pulls in favor of (2) and (3) as candidate explanations. All else being equal, we tend to prefer explanations that make the explanandum more likely. But all else is not equal. For there is another difference between these candidate explanations. (1) is far less specific than (2) or (3). And as a result, (1) is modally robust in a way that (2) and (3) are not. If (1) explains B-S, then for many nearby possible worlds where the banana rots, the explanation for the rotting would be exactly the same. Similarly, if (1) explains B-P, then for many nearby possible worlds where the vast majority of bananas rot, the explanation for the pattern of rotting will be exactly the same.

These considerations suggest that we might be able to justify P1 by first introducing the following notion:

Modal Robustness. A is a more modally robust explanation of C than B is if and only if A explains C in more nearby possible worlds than C.[19]

And then claiming that explanations are better to the extent that they are more modally robust.[20] It follows that the probabilistic explanations of statistical mechanical phenomena are better than any alternative, non-probabilistic explanations because they are more modally robust.

Something similar to this line of thinking has been used by, e.g., Weslake 2010 and Bhogal 2017 to justify taking higher-level scientific explanations to be better, in some cases, than explanations in terms of fundamental physics. But there's an important problem with using modal robustness to defend probabilistic statistical mechanical explanation over any alternative, non-probabilistic explanation. Even though (1), as an explanation of B-S and B-P, is more modally robust than (2) and (3), it isn't clear that it is more modally robust than any alternative, non-probabilistic explanation. Let M be the set that contains all and only the initial microstates for the system in question that, in combination with the fundamental laws, lead to thermodynamically normal behavior. And consider (4) as an explanation of B-S.

(4) When you left the banana on the counter for two weeks, the system started off in an initial microstate that was within set M and (in combination with the fundamental laws) microstates in M lead deterministically to rotting.

And (5) as an explanation of B-P.

(5) In most instances in which someone leaves a banana on the counter for two weeks, the system starts off in initial microstate in M and (in combination with the fundamental laws) microstates in M lead deterministically to rotting.

Here (4) and (5) are just as modally robust as (1).

[19] An interesting question here is whether we mean more worlds by number or more worlds according to some measure placed over the space of nearby possible worlds. I'll leave that question open since I don't think it impacts what follows.

[20] As Weatherson (2012) points out, it is initially implausible to claim that for any two candidate explanations A and B of explanandum C, if A is a more modally robust explanation of C than B is, then A is better explanation of C than B. Consider (1*) as an alternative explanation of B-S and B-P:

(1*) The probability of any particular banana that has been left on the counter for two weeks rotting is very high and the probability of any particular apple that has been left on the counter for two weeks rotting is very high.

(1*) is more modally robust than (1) in the sense at issue here, but (1*) does not seem like a better explanation than (1). (1*), it seems, is too unspecific.

This suggests that at best modal robustness is one of multiple criteria that factor into determining the best explanation—multiple criteria for which, insofar as they provide justification for P1, there must be a wholly objective weighting.

It is worth emphasizing here that although (4) and (5) are non-probabilistic explanations of B-S and B-P, they are not wholly *microphysical* explanations. So the notion of modal robustness proposed above might still provide a criterion by which (1) is a better explanation than any alternative, microphysical explanation—as Bhogal and Weslake suggest. The point here is just that this notion of modal robustness does not provide a criterion by which (1) is a better explanation than any alternative, *non-probabilistic* explanation. As such, an appeal to modal robustness cannot be used to justify P1. We must look elsewhere.

4.3 Counterfactual support

Here is a different line of thinking that might support taking (1), instead of any alternative, non-probabilistic explanation, to explain B-S and B-P: perhaps (1) supports counterfactuals that no alternative, non-probabilistic explanation supports.

For instance, perhaps B-C follows from (1), but not from any alternative, non-probabilistic explanation of B-S and B-P.

B-C If you had left ten seconds later, the banana would still have rotted.

Note, however, that insofar as one thinks that (1) makes B-C true, then one must have some reason for thinking so. And this reason had better not just involve a stipulation regarding the semantics of counterfactuals.

Suppose, for instance, that one stipulates that when one is evaluating a counterfactual one is allowed to ignore possible worlds in which events occur that are assigned low probability in classical statistical mechanics.[21] Then of course it will follow from (1) that B-C is true. For it follows from (1) that probability of the event of any particular banana not rotting over the course of two weeks is very low. But if the fact that B-C follows from (1) is the result merely of a stipulation about how statistical mechanical probabilities factor into the semantics for counterfactuals, then why not allow advocates of alternative, non-probabilistic explanations to make a similar stipulation? Why not allow them to stipulate that, when evaluating a counterfactual, one is allowed to ignore possible worlds in which the system starts in a microstate that is not within M (the set of initial microstates leading to

[21] This is effectively what Albert does. In Albert (2014), for instance he writes, “Find the possible world which is closest to the actual one, as measured by distance in phase-space, at the time of the antecedent, among all of those which are compatible with the past-hypothesis, and whose associated macro-histories are assigned reasonable probability-values by the statistical postulate, and in which the antecedent is satisfied, and evolve it backwards and forwards in accord with the deterministic equations of motion, and see whether it satisfies the consequent. If it does, count the counterfactual as true; if not, count the counterfactual as false” (163).

thermodynamically normal behavior)? Insofar as one ignores such worlds, B-C will follow from (4) or (5), since the worlds in which the banana does not rot over the course of two weeks are all outside M.[22]

So one cannot merely claim that probabilistic statistical mechanical explanations are better than any alternative, non-probabilistic explanations because, e.g., B-C follows from (1). Rather the claim must be that B-C follows from (1) not merely in virtue of some stipulation regarding the semantics of counterfactuals. Instead, B-C follows from (1) *due to the nature of the probabilities that show up in (1)*.

This observation puts even more pressure on Humean accounts of deterministic chance.[23] Humean accounts of deterministic chance ground the probabilities in (1) in the underlying, fully deterministic, Humean mosaic. And it is difficult to see how, given that fact, the connection between (1) and B-C could be anything more than a stipulation. There is nothing in the nature of Humean deterministic chances alone that suggests that they would support counterfactuals like B-C—or at least nothing that does not equally count in favor of (4) and (5) supporting B-C.

What kind of account of deterministic chance *would* yield the result that B-C follows from (1) in a way that does not just involve a stipulation regarding how the probabilities of statistical mechanics play a role in the semantics of counterfactuals? Consider, for instance, the view that deterministic chances are propensities. What it means to assert (1) is that bananas have a propensity to rot when left alone for two weeks. The notion of a propensity is plausibly a primitive, which doesn't admit of any sort of reductive analysis, but one can shed some light on it by pointing to nearby concepts like the concept of a tendency or a disposition. To say that bananas have a propensity to rot when left alone for two weeks is similar to saying that bananas have a tendency to rot or that they are disposed to rot under those conditions, but for the fact that propensities admit of a rigorous quantificational aspect. In particular, they obey the probability calculus.

Now tendencies and dispositions are not themselves easily understood concepts, but we do have some pre-theoretical grip on them, and that pre-theoretical grip ties them closely to counterfactuals. Part of what it means to say that A is disposed to B is that A *would* B under certain conditions. Perhaps, then, one can make the case that a propensity account of deterministic chance would yield the result

[22] Although I am making an effort to keep technical details to a minimum, it is perhaps worth emphasizing that M is not a highly disjunctive or arbitrary set. Indeed the initial microstates for any system that are not in M have measure zero on the standard Lebesgue measure that is deployed in classical statistical mechanics. All of this is to say that this sort of stipulation would not be an especially arbitrary one.

[23] Or at least it puts pressure on Humean accounts of deterministic chance insofar as such accounts are presented hand in hand with an endorsement of the explanatory role argument—remember, nothing that I am saying here is a critique of Humeanism in general; these worries are worries for Humeanism only insofar as it is endorsed in conjunction with the explanatory role argument.

that B-C follows from (1) in virtue of the nature of the probabilities in (1) themselves, not just as a result of a stipulation.

All of that is, at best, suggestive, and certainly far from conclusive. But I will leave the discussion of the second constraint that follows from the explanatory role argument there. I don't claim to have pinned down exactly what the second constraint is or how precisely it impacts one's metaphysics of deterministic chance. My primary goal has been to convince the reader that there is some such constraint. And while substantive further work needs to be done in order to spell it out, there is at least some reason to be suspicious that the Humean about deterministic chance, in particular, will be able to satisfy this criterion.

5 Conclusion

I have suggested a way of spelling out the explanatory role argument for deterministic chance in detail and shown how there are at least two ways in which this argument is likely to constrain one's metaphysics of deterministic chance. The first constraint, though fairly obvious, also raises serious concerns about Humean accounts of deterministic chance, which have been endorsed by several prominent defenders of the explanatory role argument. The second constraint is significantly more difficult to pin down, but there is good reason to think that there is such a constraint and at least one natural way of spelling it out again raises issues for Humean accounts of deterministic chance and indeed suggests a metaphysics of deterministic chance that has largely gone unexplored.

It is worth emphasizing, moreover, that the definition of Humeanism that I have relied on is very general—Humeanism, as I have presented it, is just the view that chances are "nothing over and above" the Humean mosaic. What the arguments suggest is that those who endorse the explanatory role argument cannot appeal to a familiar Humean account of how deterministic chances are related to the fundamental entities and the fundamental dynamics. And given how broadly Humeanism has been defined, the remaining options for understanding the relationship between deterministic chances and what is fundamental are all relatively underdeveloped and surprising. Perhaps, for instance, deterministic chances, despite appearing to be higher-level chances, in fact depend on some fundamental chances that arise even in worlds where the fundamental laws are deterministic. (One way to make sense of this approach is via the kind of initial chance event described in Demarest 2016.) Or perhaps deterministic chances are truly higher-level chances—but higher-level chances that are in an important sense independent of what there is at the fundamental level. Finally, it is worth noting that I have said nothing in what follows about understanding statistical mechanical claims about what is likely in terms of what is typical. (See Maudlin 2007 and Wilhelm 2022.) Those who advocate for this sort of account are usually explicit that typicality is distinct from probability, but perhaps it is possible to reinterpret the explanatory role argument such that it supports understanding deterministic chances as merely claims about what is typical. In order for this strategy to work, however, more needs to be done to determine whether explanations that appeal to typicality can be understood as metaphysically robust explanations.

What is clear at this point is that a detailed examination of the explanatory role argument has upshots both for how we think about levels of explanation in general and for how we approach the metaphysics of deterministic chance in particular. With respect to the former, the discussion above shows that those who are motivated to introduce higher-level explanations in order to respect scientific practice should be careful when pairing those explanations with substantive metaphysical commitments. Those higher-level explanations may place constraints on one's metaphysics that are surprising and unwelcome. In the specific case at hand, those who endorse the explanatory role argument for deterministic chance alongside a Humean account of the metaphysics of chance should tread carefully. It may well turn out that on their favored account, deterministic chances will end up being unfit to do the work for which they were introduced.

I would not be surprised if many readers take the upshot of this discussion to be that we should jettison the explanatory role argument for deterministic chance. But remember, that argument was motivated by scientific practice, so to jettison it would be scientifically revisionary. And one of the main constraints on contemporary metaphysics is the thought that one should not be so revisionary. Perhaps one can make the case that being scientifically revisionary in this particular case is not worrisome, but that will, presumably, require a substantive view about which aspects of scientific theorizing are up for revision and which are not.[24]

References

Albert, David Z. 2000. *Time and Chance*. Harvard University Press.

Albert, David Z. 2014. "The Sharpness of the Distinction Between the Past and the Future." In Wilson, A., and Handfield, T. (eds.), *Chance and Temporal Asymmetry*, Oxford University Press, pp. 159–174.

Beebee, Helen 2000, "The Non-governing Conception of Laws of Nature," *Philosophy and Phenomenological Research* 61: 571–194.

Bhogal, Harjit. 2017. "Special Science Explanation in a Physical World." PhD Dissertation.

Briggs, R. 2010. "The Metaphysics of Chance." *Philosophy Compass* 5/11: 938–952.

Demarest, Heather. 2016. "The Universe Had One Chance." *Philosophy of Science* 83 (2): 248–264.

Emery, Nina. 2015. "Chance, Possibility, and Explanation." *British Journal for the Philosophy of Science* 66 (1): 95–120.

[24] Many thanks to Katie Robertson and Alastair Wilson for helpful feedback. Thanks also to Harjit Bhogal, Heather Demarest, Foad Dizadji-Bahmani, Barry Loewer, Sam Mitchell, Elizabeth Miller, Isaac Wilhelm, and audiences at the 2018 CHPS conference in Boulder, at the Foundations of Probability Seminar at Rutgers and at the Chance Encounter Workshop at the University of Gröningen.

Emery, Nina. 2017. "A Naturalist's Guide to Objective Chance." *Philosophy of Science* 84 (3): 480–499.

Emery, Nina. 2018. "Laws and Their Instances." *Philosophical Studies* 176 (6): 1535–1561.

Emery, Nina. 2020. "Laws of Nature." In Raven, M. J. (ed.), *Routledge Handbook of Metaphysical Grounding*, 437–448, Routledge.

Emery, Nina. 2023. "The Governing Conception of Laws." *Ergo* 9.

Feynman, Richard, Robert Leighton, and Matthew Sands. 1963. *The Feynman Lectures in Physics Vol. II*. Pearson/Addison-Wesley.

Glynn, Luke. 2010. "Deterministic Chance." *British Journal for the Philosophy of Science* 61 (1): 51–80.

Hájek, Alan. 1996. "'Mises Redux'—Redux: Fifteen Arguments against Finite Frequentism." *Erkenntnis* 45 (2–3): 209–227.

Handfield, T., and Wilson, A. 2014. "Chance and Context." In Wilson, A., and Handfield, T. (eds.), *Chance and Temporal Asymmetry*, Oxford University Press, pp. 159–174.

Hicks, Michael T., and Peter van Elswyk. 2014. "Humean Laws and Circular Explanation." *Philosophical Studies* 172 (2): 433–443.

Hicks, Michael T., and Alastair Wilson. 2021. "How Chance Explains." *Noûs* 57 (2): 290–315.

Hoefer, Carl. 2007. "The Third Way on Objective Probability: A Sceptic's Guide to Objective Chance." *Mind* 116 (463): 549–596.

Ismael, J. T. 2009. "Probability in Deterministic Physics." *Journal of Philosophy* 106 (2): 89–108.

Lange, Mark. 2013. "Grounding, Scientific Explanation, and Humean Laws." *Philosophical Studies* 164 (1): 255–261.

Lewis, D. 1986a. "A Subjectivist's Guide to Objective Chance." *Philosophical Papers* 2: 83–132. Oxford University Press.

Lewis, D. 1986b. *Philosophical Papers Volume II*. Oxford: Oxford University Press.

Loewer, Barry. 2001. "Determinism and Chance." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32 (4): 609–620.

Loewer, Barry. 2012. "Two Accounts of Laws and Time." *Philosophical Studies* 160 (1): 115–137.

Maudlin, Tim. 2007. "What Could Be Objective about Probabilities?" *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 38 (2): 275–291.

Meacham, Christopher J. G. 2005. "Three Proposals Regarding a Theory of Chance." *Philosophical Perspectives* 19 (1): 281–307.

Meacham, Christopher J. G. 2010. "Two Mistakes Regarding the Principal Principle." *The British Journal for the Philosophy of Science* 61 (2): 407–431.

Miller, Elizabeth. 2015. "Humean Scientific Explanation." *Philosophical Studies* 172: 1311–1332.

Schaffer, Jonathan. 2007. "Deterministic Chance?" *British Journal for the Philosophy of Science* 58 (2): 113–140.

Schaffer, Jonathan. 2008. "Causation and Laws of Nature: Reductionism." In Hawthorne, J., Sider, T., and Zimmerman, D. (eds.), *Contemporary Debates in Metaphysics*, 82–107. Oxford: Basil Blackwell.

Strevens, Michael. 2000. "Do Large Probabilities Explain Better?" *Philosophy of Science* 67: 366–390.

Wallace, D. 2011. "The Logic of the Past Hypothesis." *PhilSci Archive*, University of Pittsburgh. https://philsci-archive.pitt.edu/8894/

Weslake, Brad. 2010. "Explanatory Depth" *Philosophy of Science* 77 (2): 273–294.

Weatherson, Brian. 2012. "Explanation, Idealisation, and the Goldilocks Problem." *Philosophy and Phenomenological Research* 84: 461–473.

Wilhelm, Isaac. 2022. "Typical: A Theory of Typicality and Typicality Explanation." *British Journal for the Philosophy of Science* 73 (2): 561–581.

Wilson, Jessica. 2014. "No Work for a Theory of Grounding." *Inquiry* 57 (5–6): 535–579.

PART VI

HOW ARE EXPLANATORY LEVELS POSSIBLE?

16
Why Are There High-Level Regularities?

Harjit Bhogal

> So, then, why is there anything except physics? That, I think, is what is really bugging Kim. Well, I admit that I don't know why. I don't even know how to think about why. I expect to figure out why there is anything except physics the day before I figure out why there is anything at all, another (and, presumably, related) metaphysical conundrum that I find perplexing. (Fodor 1997, 161)

Fodor is puzzled about why there is anything except physics. His specific puzzlement here isn't about why there exists morality or aesthetics, or all sorts of other things that aren't physics—though clearly those are interesting puzzles too. Rather, his puzzlement is about why there are special sciences. Why, in addition to physics, do we have biology, economics, sociology, geology, and so on? This is particularly puzzling if we assume, as I do, that our world is ultimately physical—that the basic constituents of the world are physical, and all of the higher-level features of the world are made up of these constituents.

This puzzle can be approached in various ways. One way is to investigate what advantage there might be to studying the world in special scientific terms rather than physical terms—to find some reason why we would bother to do biology, for example, in addition to physics. There's lots to say here. For example, there is a large literature on how explanations given by these special sciences might be superior to the lower-level explanations given by physics.[1] If special science explanations really are superior, at least some of the time, then this is one reason to pursue special sciences as well as physics. Another natural thought is that there is a pragmatic advantage to studying the world at the higher level. It's just too hard to do the physics of interest rates, for example—economic investigation is more tractable.

But prior to the question of what makes the special sciences worth pursuing is the question of why there exists anything that can be reasonably pursued.

[1] Garfinkel (1981), Jackson and Pettit (1992), Hitchcock and Woodward (2003), and Strevens (2008) are influential examples of different approaches to explaining the advantages that higher-level explanations can have.

Harjit Bhogal, *Why Are There High-Level Regularities?* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Harjit Bhogal 2024. DOI: 10.1093/oso/9780192862945.003.0017

Why, that is, do the basic requirements for the existence of higher-level sciences hold?

What requirements are these? For there to be higher-level sciences there presumably have to be higher-level counterfactuals and causation and explanations, and so on. And, again, there are literatures on all these issues. But the most obvious requirement is that there must be high-level regularities. Such regularities don't need to be exceptionless—most regularities in the special sciences seem to be exception-tolerant in one way or another. But there must, at least, be *some* patterns, *some* high-level regularities. You couldn't have a discipline of economics, for example, without there being some general things to say about the allocation of resources.[2]

That's the point I will focus on—about the existence of such high-level regularities. It's the point that Fodor was focused on too—he glosses the question "why is there anything except physics?" as the question of "why there should be (how there could be) macrolevel regularities at all in a world where, by common consent, macrolevel stabilities have to supervene on a buzzing, blooming confusion of microlevel interactions" (Fodor 1997, 161).

That's what I'm going to consider—why are there high-level regularities when, at the basic level, things are just a blooming, buzzing confusion? Isn't it baffling, for example, that the almost unimaginable array of fundamental particles and fields that make up an economy somehow choreograph themselves into the regularity that the average return on a stock is inversely proportional to its covariance with the market? And isn't it even more baffling that this isn't even close to being an isolated case? We have sociology, ecology, development economics, geology, biochemistry, neuroscience, microeconomics, meteorology, immunology, oceanography, thermodynamics, and many many more special sciences, all with their proprietary regularities. And aren't such regularities *even more* baffling when we recognize, as Fodor (1997) stresses in response to Kim (1992), that these regularities connect kinds that are physically heterogeneous. Stocks, and their prices, for example, can be realized by systems that are vastly physically different—the ticker-tape-driven stock market of the 1920s is very physically different from the algorithm-driven stock market of today.

[2] Perhaps this is a little quick. Maybe there are some disciplines that proceed not by way of generalization, but by detailed investigation into specifics. Certain types of anthropology might, at least on some conceptions, be disciplines of this kind. And there is a long tradition of theorizing about social science that, in one way or another, points to there being a distinction between disciplines, or at least methodologies, that generalize and those that specify. (Consider, for example, the distinctions between *nomothetic* and *idiographic* (Windelband 1904) approaches to enquiry and the distinction between *erklären* and *verstehen* as different types of explanation and understanding (e.g., Dilthey 1894)).

Some will be inclined to respond that disciplines that don't generalize don't count as sciences, so all *special sciences* require regularities. But regardless of whether we agree with that, there is clearly an important set of special sciences for which the existence of high-level regularities is a prerequisite.

Is it just a coincidence, a miraculous piece of luck, that the blooming, buzzing confusion of the lower level gives rise to such high-level stability? Or is there something more substantial we can say?

In this chapter I will, in a rather abstract and schematic way, consider a few different strategies for explaining the existence of high-level regularities, ending with a novel strategy of my own. The key idea of this strategy is that we can transform the question of high-level regularities in a way that makes accounts of *natural properties* important for answering the question. Then we can appeal to certain accounts of natural properties in the special sciences to answer the transformed question.

The aim is to understand the landscape of strategies for explaining the existence of higher-level regularities rather than on detailed implementation. As such my discussion of certain strategies, including my own favored strategy, will be rather quick—there is a huge amount of detail that is missing.

The main takeaway from looking at the landscape from this height is that a certain kind of "bottom-up" strategy for explaining why there are high-level regularities—one which involves detailed consideration of our basic science—is not fully satisfying. This encourages us to, in addition, consider some "top-down" strategies—ones which require attending to metaphysical detail more than scientific detail. These more metaphysical strategies are certainly not in conflict with the more scientifically oriented, bottom-up strategies, but they address slightly different questions. In particular, Fodor's question is fruitfully approached in this more metaphysical way.

So, let's consider some possible responses to this puzzle about special science regularities.

1 Fodor

Fodor's reaction to this puzzle was to be baffled, but in an interestingly specific way. Regarding why special science regularities exist he says: "Well, I admit that I don't know why. I don't even know how to *think about* why. I expect to figure out why there is anything except physics the day before I figure out why there is anything at all, another (and presumably related) metaphysical conundrum that I find perplexing" (Fodor 1997, 161, emphasis in original).

This type of bafflement suggests an interesting position. In the rest of the section I will discuss this position. I'll write as if it was Fodor's position, and I think a decent case can be made that it was. But, ultimately, I don't care about Fodor exegesis—I'm merely interested in a position that is suggested by some things that Fodor says.

There's a natural interpretation of Fodor (1974) which takes him to deny physicalism, at least on some reasonable conceptions of physicalism (Loewer (2009) makes this case in detail). For example, he says that:

> I am suggesting, roughly, that there are special sciences not because of the nature of our epistemic relation to the world, but because of the way the world is put together: not all natural kinds … are, or correspond to, physical natural kinds. (Fodor 1974, 113)

The suggestion seems to be that some features of the world don't hold in virtue of the physics. There are facts about special science natural kinds, for example, that are not fixed by physics. As Loewer (2009) notes, Fodor says similar things with respect to special science laws—they don't seem to hold in virtue of the physical facts on his view. If there are things in the world that don't hold in virtue of the physics, then there is a natural sense in which physicalism is false. (Of course, there are other reasonable senses of physicalism, but getting into that debate would take us off topic.)

If special science laws and kinds are just a basic part of the way that the world is put together, then Fodor's specific bafflement makes sense. He claimed that the question of why there is anything except physics was similarly puzzling, and related to the question of why there is something rather than nothing. The question of why there is something rather than nothing seems so intractable because it's about the fundamental makeup of the world. Of course, things exist now because of entities that existed in the past. But, why were there any such entities? This is a question about the world's fundamental starting points. And part of what it is to be fundamental is for there to be nothing that is explanatorily prior. So there seems to be no material we can use to give an explanation.

If we think that special science laws are part of the fundamental makeup of the world, as Fodor seems to do, then the question of why there exist special science regularities seems intractable in the same way.

But maybe this isn't so bad. There is a sense in which this position might defuse the worry about the existence of the special sciences, or at least make it less pressing. If you think that special science laws are basic, then it might seem reasonable to say "that's just the way the world is put together." Because what else can we say about the fundamental? This position suggests that there is no special puzzle about special science regularities—no more than the puzzle about why there are physical regularities.

This approach is, I think, interesting and not something that should be rejected out of hand. But I'll just quickly raise three concerns. Firstly, there is obviously a sense in which this isn't a satisfying explanation of the puzzle. It is, at best, a reason to think that we shouldn't look for an explanation. This is a reasonable fallback position, but it would be better to have a genuine explanation. Secondly, the view is not physicalist, in the sense that there are parts of the world that don't hold in virtue of the physics. This seems deeply unattractive to me. But, regardless of that, the puzzle that I started out with, that I want to address in this paper, is the puzzle of why there is anything except physics, when, in a sense, physics is all that there is. So for the rest of the paper I'm going to consider positions which accept that the world is ultimately physical. And thirdly, the view that certain special science laws are just part of the way that the world is put together—that such laws are fundamental—leads to concerns of redundancy. If the fundamental physical laws already make all the physical events happen, and if the higher-level entities are made up out of physical entities, then it can be hard to see what the fundamental special science laws are there to do. (Loewer (2009) develops this kind of objection to Fodor. And, of course, this point is very closely related to huge literature on mental causation and overdetermination.)

None of these concerns rule out this Fodorian position, but they motivate us to look elsewhere.

2 Anthropic Reasoning

Here's an approach to the puzzle that we can put aside quickly. We might attempt to use anthropic reasoning to explain the existence of special science regularities. The basic thought is simple entities like us would not exist without there being some higher-level regularities. If there were no regularities about the working of DNA, for example, then we would not exist. So, it is not surprising that we live in a world where there are such regularities.

Such anthropic reasoning is highly controversial. It's unclear whether such reasoning really can generate good explanations, and under what conditions. But, regardless of these complicated questions, it can't be a satisfying answer to our puzzle because most high-level regularities are not required for our existence. The regularities of microeconomics, for example, are not required for our existence. Similarly with the regularities of ecology, sociology, meteorology, psychology, and so on. So we should look for a more general story about special science regularities. In fact, the idea that what is needed is a "general" story will be important in what's to come.

3 Using the Physics to Explain

3.1 Metaphysical Explanation

Here is perhaps the most obvious thought about how to explain the existence of special science regularities—we are assuming that all the special science facts hold in virtue of physical facts, so it must be that the special sciences regularities are explained by the physics.

There is a sense in which this is correct—the special science regularities are *metaphysically* explained by, or *grounded in*, the physical facts. So the fact that average returns on stocks are inversely proportional to their covariance with the market is grounded in the facts about the fundamental particles and fields and so on.

But this type of metaphysical explanation is not what we are looking for. In fact, such an explanation is part of the puzzle—given that the special sciences are grounded in the buzzing, blooming, confusion of fundamental physics, how do higher-level regularities arise? The mere fact that there are lower-level realizers of special science regularities doesn't make it unsurprising or non-coincidental that there are such regularities.

3.2 Precise Scientific Explanations

So we can put aside such metaphysical explanations. But still, it might seem that the assumption of physicalism guarantees that there is a kind of scientific explanation of special science regularities. For simplicity, assume that the physical laws are deterministic. Then we might explain the existence of higher-level regularities by appealing to the physical laws and the precise initial conditions of the universe, showing the world evolves into a situation where there are high-level regularities.

Again this is clearly an unsatisfying explanation. The mere fact that there are initial conditions and laws that led to there being lots of special science regularities doesn't make it unsurprising or non-coincidental that there are such regularities.

The reason this explanation is unsatisfying is because it is extremely *fragile*. It says that the reason special science regularities exist is the precise microphysical details of the initial conditions and the precise laws. But, of course, those conditions could extremely easily have been different.[3]

If all we can say about why there are special science regularities is that they hold because of the precise initial conditions and laws, then it is just a fluke or coincidence that there are such regularities. Analogously, imagine that we toss a coin and it lands heads 100 times in a row. We could explain this by citing the exact initial conditions of the universe and the deterministic laws, but if that's all we can say, then it's still just a fluke or coincidence that the coin landed heads every time.

The appeal to precise physical details doesn't give us the satisfying explanations that we want.

[3] The precise way in which such explanations are unsatisfying is investigated in great detail in the literature on levels of explanation. I framed the issue in terms of modal fragility—in the spirit of Wilson (1994), Weslake (2010), Bhogal (2020b), and others. But others think that the problem is something subtly different, for example, that the explanation doesn't cite the proper *difference-makers* (e.g., Strevens 2008); or the explanans fails to be *proportional* to the explanandum (e.g., Yablo 1992; Woodward 2001). These are all fairly closely related. The differences don't matter for the purposes of this paper.

3.3 Generic Scientific Explanation

Explanations which appeal to the precise physical details aren't what we want. Perhaps we can do better by appealing to more generic features of the physics. Such explanations can be more satisfying. Imagine that instead of appealing to the precise physics to explain the 100 heads coin tosses we cite the fact that the coin is weighted in such a way to make it overwhelmingly likely to land heads. Clearly this would show the regularity to be non-coincidental.

How can we give such an explanation of high-level regularities though? Perhaps the most influential idea, in the recent literature, stems from Batterman (2000) and his discussion of *renormalization group* (RG) explanations.

The literature on RG explanations is technical and complicated, and here is not the place to get into the physical details. But the basic idea is to start with a particular physical system and a function, specifically a Hamiltonian, that describes that physical system. Then we apply a particular transformation to the Hamiltonian to output another Hamiltonian that intuitively represents the system at a larger scale. Then we can apply the same transformation again to the resulting Hamiltonian. Sometimes, when we repeatedly apply this transformation to the Hamiltonians that represent different physical systems, we find that the transformed Hamiltonian ends up being the same for both systems. This tells us that at a certain scale, or level of grain, the different systems exhibit the same behavior.

This procedure allows for a kind of explanation of high-level regularities. If we can show that there is a large class of physical systems, F, such that when we repeatedly apply the transformation their Hamiltonians all flow into the same transformed Hamiltonian, then we know that all of those physical systems realize the same high-level regularity. In doing this it seems like we can explain the high-level regularity by identifying the generic features of physical systems that really make a difference to the holding of the high-level regularity—the features that make a physical system part of the class, F. This explanation is much more satisfying than simply appealing to the precise details of the physics.

Notably, though, this approach is extremely piecemeal. It has to be applied to every high-level regularity, one at a time—trying to find the features of physical systems that make a difference to the holding of that regularity. And although that's an extremely interesting scientific project, it has a lot of limitations for trying to explain why the world has special science regularities.

Firstly, the project of trying to explain all special science regularities in this way is rather optimistic. When we consider the huge number and diversity of special science regularities—again, we have sociology, ecology, development economics, geology, biochemistry, neuroscience, microeconomics, meteorology, immunology, oceanography, thermodynamics, and so on, all with distinct regularities—then executing this piecemeal approach, taking one regularity at a time, starts to seem extremely intimidating.

Moreover, it seems like RG strategies will not work for the vast majority of special science regularities. RG explanations have only been worked out for very few cases, notably the similar behavior of strikingly different systems during phase transitions. And the prospect for giving renormalization group explanations of, for example, the regularities expressed by the Lotke-Volterra equations in population ecology seem slim at best.

Batterman (2000) accepts that there are limits to the applicability of renormalization group explanations. For example, citing Block (1997, 120), he asks "how can considerations of structural stability play a role in explaining how an 'and'-gate can be instanced in silicon, in hydraulics, and in cats, mice and cheese?" Batterman notes that RG explanations may not be appropriate for such cases—it doesn't seem like we can use RG explanations to explain regularities about "and"-gates. He accepts that it's unlikely that we could use the mechanics of RG explanations to explain all high-level regularities (116–117).

So it looks like the RG strategy can't apply to a wide enough range of cases to resolve our puzzle about high-level regularities. Rather, he hopes, we will use other explanatory strategies to the same end, the end of identifying the generic physical facts that make a difference to the holding of particular high-level regularities. But, currently, we have little idea what such explanatory strategies would look like.

But this concern about optimism is not the key point. More important is that, even if we could execute this explanatory strategy for a wide range of special science regularities, there is still a sense in which it doesn't satisfyingly answer the question of why there are special science regularities.

Again, the strategy we are considering is a piecemeal one. To execute this properly would consist in going through the special science regularities we have, one by one, and finding, for each one, the physical features that give rise to the regularity. This would explain specific regularities, but it isn't really a satisfying explanation of why there are special science regularities at all. Consider, again, trying to explain why the coin we tossed 100 times lands heads every time. We could take a piecemeal approach and attempt to explain the regularity by separately explaining why each coin toss landed heads. That is, we could try to explain why the first coin landed heads by appealing to the way it was tossed and the physical laws, and we could do the same for the other 99 tosses. Even if we were successful in doing this, simply separately explaining each toss doesn't satisfyingly explain why all the coin tosses landed heads. In a similar way, conjoining lots of separate explanations of specific special science regularities doesn't satisfyingly explain why we are in a world which contains a mass of high-level regularities. It doesn't bring us closer to seeing how the buzzing, blooming confusion of the physical level leads to regularity.[4]

[4] There are some cases where this strategy could result in a good explanation of why there are higher-level regularities. Perhaps if it turned out that when we separately explain all the higher-level

In general, bottom-up approaches, like the appeal to RG methods, address the question of why we have the particular special science regularities that we do. But this, I think, wasn't what Fodor was puzzled about. Fodor was puzzled about why there are special science regularities *at all*. Fodor's bafflement would not be resolved by pointing out to him, as if he didn't know, that there are some scientific strategies for explaining why certain higher-level regularities hold in terms of lower-level science and that hopefully such strategies could be found for all special science regularities.

Explaining why there are regularities and explaining the particular regularities that hold are different projects. Of course, if we successfully explain the particular regularities that hold, then that explanation will entail that there are regularities. The fact that there are special science regularities is realized by the particular regularities that hold. But it's a familiar point to philosophers of science, particularly those who think about levels of explanation, that an explanation of a realizer of a fact is not, in general, a satisfying explanation of the fact itself—particularly when the fact can be realized in many different ways.

For example, consider Kitcher's (2001) discussion of *Arbuthnot's regularity*—that each year between 1623 and 1705 more boys were born than girls in London. Kitcher notes that we could employ the following strategy to explain the regularity:

> Start with the first year (1623); elaborate the physicochemical details of the first copulation-followed-by-pregnancy showing how it resulted in a child of a particular sex; continue in the same fashion for each pertinent pregnancy; add up the totals for male births and female births and compute the difference. It has now been shown why the first year was "male"; continue for all subsequent years. (71)

But, he claims, even if we could do this "it would not show that Arbuthnot's regularity was anything more than a gigantic coincidence." Explaining the particular realizer of Arbuthnot's regularity does not satisfyingly explain the regularity.[5]

Similarly, piecemeal explanation of the particular special science regularities is extremely scientifically valuable, but it doesn't get us everything we would want. It leaves the general fact that there are regularities at all without a satisfying explanation.

regularities we find that it's the same features of the physical level that are the difference-makers for the holding of all the regularities, then that might bring us close to an explanation of why we live in a world with such a mass of high-level regularities—it's because those physical-level features hold. But we have don't currently have reason to expect that such a situation would result.

[5] In Bhogal (2020a) I give an account of coincidence based on these considerations.

4 A Top-Down Approach

These considerations suggest that we need a different type of strategy for explaining why there are special science regularities. But if we are to explain why there are special science regularities and not merely explain the particular regularities that hold, then the strategy will have to be less tied to the scientific detail. The strategy will have to be more "metaphysical."

4.1 Transforming the Question

I'm going to sketch such a strategy. It's one that is top down and general rather than bottom up and piecemeal. Though, given space constraints, I can only outline the strategy. The aim is just to see what a strategy that is aimed at explaining the existence of regularities in general would look like.

The first part of the approach involves transforming the question of why there are high-level regularities into something that we can get more of a grip on. And the first step in this transformation is to emphasize something that is familiar to metaphysicians but sometimes gets ignored by philosophers of science. Properties, at least on one conception, are *abundant*. As well as mass, charge, and greenness, there are properties like *grue*.

Grue is fairly simple to define—we do it with few problems in our undergraduate classes. But there are many, many other properties that we can characterize via ridiculously complicated definitions with millions of conjunctions and billions of disjunctions. There are, clearly, *a lot* of properties. (Very plausibly, there is an uncountable infinity of just length properties—being 1 meter long, being π meters long, etc.).

This abundance of properties is relevant because it makes it, in one sense, extremely unsurprising that there are so many high-level regularities. There are just *so many* properties that there will be simple universal generalizations that we can state, regardless of what the world is actually like. For example, Lewis (1983, 367), when discussing his view of laws of nature, considers a predicate F that holds of all and only the objects in the actual world. *Everything is F* will, then, be a very simple regularity that holds in our world. It's easy to see how to generate other high-level regularities in similar ways. There will be a property G such that everything in the USA is G. There will be properties H and J such that every time a stock has property H then it had property J yesterday. If we really make use of the idea that we can define up predicates in a vast variety of different ways, with unlimited complexity, then it's easy to see that there will be many, many regularities, no matter what the world is like.

But, of course, this isn't a satisfying answer to the question of why there are high-level regularities. *Everything is F* isn't what we were thinking of when we asked why the world contains higher-level regularities. The reason we were asking that question was to understand why the special sciences exist, and regularities like *everything is F* are not the type that science investigates.

But why don't the special sciences investigate such regularities? Clearly the problem is with the property F—it's strangely *gerrymandered* and not *projectable*. To use the terminology most common in the metaphysics literature, property F seems *unnatural*. Let us take a moment to consider the concept of naturalness in a little more detail.

4.1.1 Naturalness

It's common for metaphysicians to recognize the need for a distinction between *natural* and *unnatural* properties. The literature here largely stems from Lewis (1983). His key idea was that natural properties play a variety of important theoretical roles, to do with laws, similarity, induction, reference, causation, and so on. Scientific laws are about natural properties; sharing of natural properties is what makes for similarity between objects, and so on.

(There is another common sense of naturalness—the naturalness of *natural kinds*. Although there are relations between natural properties and natural kinds, they are importantly different. Much of the literature on natural kinds revolves around issues of classification—how to classify organisms into species, for example. This is somewhat different from what is going on in the natural properties literature, where we are looking for a set of properties that play important theoretical roles—roles to do with induction, causation, explanation, and so on.

This is not to say that the work that goes on under the heading "natural kinds" ignores the connections between natural kinds and laws, causation, or explanation. But, nevertheless, we can consider two distinct projects—one about properties that play a special theoretical role and the other about natural groupings. It's this former project that is relevant for our current discussion of why regularities like *everything is F* aren't investigated by the special sciences.)

The literature that has followed Lewis has largely focused on fundamental physical properties and on a primitivist conception of naturalness. The natural properties (or, at least, the perfectly natural properties) were taken to be things like *spin* and *charge*, and their naturalness taken to be a basic, irreducible fact. Call such fundamental physical natural properties *F-natural* properties.

But just as there is reason to accept F-natural properties, there is similar reason to accept a distinction between natural and unnatural *special science* properties. Just as there are certain, natural properties that play central roles in the practice of physics, so there are certain natural special science properties which play analogous roles with respect to the practice of the special sciences.

For example, it seems like there cannot be special science laws about unnatural properties. Fodor (1974, 102) makes this point when he says: "I take it that there is no natural law which applies to events in virtue of their being instantiations of the property *is transported to a distance of less than three miles from the Eiffel Tower*." Similarly, such unnatural properties are not good candidates for performing inductions on. Neither do they seem to be good candidates for giving explanations of other facts we care about.

That is all to say that it seems like some higher-level properties are unnatural, and because of this, our special science theorizing should not be framed in terms of those properties—such properties don't play the right roles with respect to laws, induction, and explanation.

Regularities like *everything is F* aren't investigated by the special sciences because F is unnatural—it doesn't play the relevant roles in our special science theorizing. Even though the abundance of properties guarantees the existence of regularities like *everything is F*, that doesn't solve our problem. When we are asking why there are high-level regularities, we are not asking about the totality of properties—that would make the answer too easy. We are only asking about the natural properties—the ones that can play the relevant roles in science. So the question we are really interested in is why are there high-level regularities about relatively natural properties?

Now we have transformed the question there is a new way of attacking it—the concept of naturalness gives us a hook. Perhaps we can answer the question about high-level regularities by giving a story about natural properties.

5 Regularities and Accounts of Special Science Naturalness

The first step was to transform the question to see the relevance of accounts of special science naturalness. The second step is to give an account of special science naturalness that will help us answer the question.

Clearly this strategy is driven by the metaphysics more than the scientific detail. But, as we noted, that is appropriate if we are trying to explain why there are special science regularities and not just explaining the regularities there are.

I certainly don't have space to fully develop an account of special science naturalness. And there aren't many developed accounts of special science naturalness in the literature to lean on. So I'll quickly mention a couple of accounts that don't seem to help us with the question at issue before looking at the structure of a couple of accounts that might do the job.

As we noted in Section 4.1, the most common approach to the naturalness of the fundamental properties is to take their naturalness as a primitive. Similarly, you might take a primitivist view of special science naturalness. Clearly this view doesn't help with our question. That some high-level properties have this primitive feature of naturalness doesn't explain why there are regularities about such properties. In fact, you might think, it just adds another mystery. Not only do high-level regularities arise from the buzzing, blooming confusion, but, miraculously, the properties involved in those regularities have this primitive feature of naturalness.

Perhaps the most common approach to special science naturalness is to derive it from the primitive naturalness had by the fundamental properties—what I've been calling F-naturalness (Lewis (1986, 62), Sider (2011, section 7.3)). The basic idea is to give an account of graded F-naturalness. The graded F-naturalness of a property is fixed by the length of that property's definition in terms of the perfectly natural, fundamental properties. Shorter definitions make properties more F-natural.[6] The special science natural properties, on this approach, are supposed to have a high degree of F-naturalness, or at least a higher degree than intuitively unnatural properties.

This account might give us hope for explaining the existence of high-level regularities. If we assume that there are regularities about the F-natural properties, then we *might* be able to argue that there will also be regularities about properties that are fairly simply defined in terms of those properties. Since the special science natural properties are ones that have shorter definitions in terms of the fundamental properties, *maybe* this helps us see why there are regularities about those properties.

But this hope dissipates when we focus on just how ridiculously long and complicated the definitions of the special science properties like *inflation* would be in terms of the fundamental properties. There is no hope, I think, of leveraging the existence of regularities about the fundamental properties into an explanation of the existence of regularities about high-level natural properties.

So what would an account of special science naturalness have to look like in order to help us answer the question? Well, if we have a reductive account of special science naturalness where part of what it is for a property to be natural is for it to be integrated with special science theorizing in the right way, then we might be able to explain why there are regularities about the high-level sciences. To put it in the bluntest way possible, if part of what it is for properties to be natural is for there to be regularities about those properties, then it's not surprising that there are regularities about the natural properties.

Here's another way to put it. Remember, there is a vast abundance of properties. An account of special science natural properties will narrow down this abundance to a special set of properties. And in this vast abundance of properties there is a set of properties that have fairly simple regularities about them. So, this invites an account of special science naturalness where the properties that we identify as natural are a subset of the properties that have simple regularities.

[6] Perhaps additional factors are relevant for degree of F-naturalness. For example, perhaps definitions that involve lots of disjunctions make for more unnatural properties than those involving lots of conjunctions, even when the definitions are the same length (see Dorr and Hawthorne (2013)).

The idea that the naturalness of a property is closely related to there being regularities about those properties is suggested by Lewis.

> Thus my account explains … why the scientific investigation of laws and of natural properties is a package deal; why physicists posit natural properties such as the quark colours in order to posit the laws in which those properties figure, so that laws and natural properties get discovered together. (Lewis 1983, 368)

A large part of the reason that we think certain properties are natural is because we can state regularities about them and theorize effectively using those properties and regularities. Lewis was making this point about the *epistemology* of naturalness.

But if we are disinclined toward primitivism about special science naturalness then the idea that a property being natural is closely tied to its scientific role in formulating interesting regularities seems like a good place to start in giving a metaphysical story.

So, I'm going to point toward a couple of accounts of special science naturalness that are of this form. Again, space constraints mean that the details of these accounts will have to be left for elsewhere.

5.1 The Package Deal Account

The first such account is directly inspired by the Lewis quote above—it's the *package deal account* (PDA) of laws and natural properties that has been developed by Barry Loewer in a series of papers and a forthcoming book (1996, 2007, 2020, forthcoming).

This account is based on the Best System Account (BSA) of laws of nature. The basic idea of the BSA is that the laws are propositions that are relatively simple, but also informative about the mosaic of occurrent facts. More precisely, consider sets of axioms. Some sets of axioms are informative about the mosaic—their deductive closure tells us a lot about the mosaics. Some sets of axioms are simple, in the sense of being syntactically simple when written down. The laws are the set of axioms that best balance simplicity and informativeness.

The PDA aims to adapt the BSA so that it outputs the natural properties, as well as the laws. Roughly speaking, the BSA says that the laws are the propositions that are simple and informative about the mosaic of occurrent facts. On the PDA view, roughly speaking, the natural properties are the properties that are referred to in those simple and informative propositions (perhaps with some other conditions added).

(The BSA has traditionally been developed with a focus on the laws of fundamental physics. Consequently, a PDA that is based on such an account of laws will not output special science natural properties. But, there are many suggestions for how to adapt the BSA in order to capture special science laws (e.g., Schrenk 2006; Albert 2000; Loewer 2001). A version of the PDA that aims to output special science natural properties should be built upon such an adapted account.)

This account makes it easy to see why there are high-level regularites about such natural properties—what it is for properties to be natural is, in part, for there to be sufficiently simple and informative regularities about them.[7] There is a huge amount more to say about the PDA. Properly developing the account is extremely complicated, and it's not clear whether it can succeed. (I discuss some of these issues in Bhogal (2023).) But it's clear how an account of this form can explain the existence of high-level regularities.

5.2 Explanatory Clusters

Here's another account of special science naturalness that might help us, one that I've developed in other work (Bhogal ms.). The basic idea is that the special science natural properties are those that form *explanatory clusters*. Roughly speaking, a set of properties forms an explanatory cluster when most of the facts about those properties are explained well by other facts about those properties.

Consider, for example, microeconomic properties. Facts about demand for goods are explained by agents' preferences; facts about certain preferences are explained by other preferences; facts about certain choices are explained by preferences and prices; facts about prices are explained by facts about demand and supply; facts about the existence of certain goods are explained by the demand for other goods; facts about the supply of goods are explained by the demand for certain factors of production; facts about the demand for factors of production are explained by the price of the goods that they are used in producing; and so on.

What we have here is a cluster of properties connected by robust explanatory patterns. The basic microeconomic properties are deeply connected and integrated. For another example, consider classical genetics. *Gene*, *allele*, *trait*, *dominance*, and *inheritance* are all closely connected by good explanations and will form a cluster. Further, consider population ecology and properties like *population*, *generation*, *predator*, *prey*, *carrying capacity*, and thermodynamics and properties like *temperature*, *pressure*, *entropy*. In general, successful special sciences seem to come with such clusters of explanatory properties—the basic properties in terms of which explanatory theories in those domains are formulated.

[7] Loewer, in personal correspondence, has also suggested that the PDA could explain the existence of regularities.

Again, there is a huge amount more to say about this strategy. Saying precisely when properties form an explanatory cluster and when they do not takes a lot of work. And arguing that this account will be extensionally adequate takes even more. Those are tasks for elsewhere.

But we can see how such an account would help answer our question. What it is for a property to be natural is for it to be part of robust explanatory patterns that connect it to other properties. And those explanatory patterns will be high-level regularities. For example, in microeconomics, one such robust explanatory pattern is that facts about demand for goods are explained by agents' preferences. This explanatory pattern generates high-level regularities connecting demand and preference. When properties form explanatory clusters, high-level regularities will result.

Furthermore, it should be unsurprising that there are such explanatory clusters. As we stressed in Section 4.1, there is such an abundance of properties—such a huge infinity of properties that we can define up in arbitrarily complicated ways—that it's deeply unsurprising that some will be clustered in the relevant way.

I'm not making the claim that it's *guaranteed* that there will be such clusters—maybe there are some possible worlds where there are no clusters to be found. But when we consider just how easy it is to define up properties it starts to seem rather unlikely that there are no sets of properties that are closely connected in the way that, for example, preference and demand are connected.

This account of naturalness helps answer the question of why there are regularities about special science natural properties in roughly the same way as the PDA account. On both the PDA and the explanatory clustering approach, properties count as natural if they play certain important roles in our scientific theorizing. That there are properties that play such roles is made very likely by the vast number of properties that exist.

The PDA approach is about properties being part of simple ways of summarizing the world. The explanatory clustering approach is about those properties being part of rich explanatory networks. But both of these roles in scientific theorizing imply that there are regularities about the relevant properties.

Further, notice that, on both these accounts, what it is to be natural isn't some obscure thing. Some philosophers of science look upon the concept of naturalness with suspicion—as a strange postulate where our judgments about it are driven by bias and presupposition.[8] But, on these accounts, what it is for a property to be natural is integrated with the practice of scientific theorizing—with finding informative generalizations and drawing explanatory connections. There's little cause for similar suspicion about naturalness in this sense.

[8] Thanks to a reviewer here.

6 Objections

So that's the strategy. It's extremely simple. First, you transform the question of why there are high-level regularities to the question of why there are high-level regularities about natural properties. And second, you give an account of natural properties which builds in there being regularities about those properties. That there are properties that meet the conditions of this account is made likely by the abundance of properties there are.

There are, of course, reasons someone might be doubtful. Obviously, it's reasonable to doubt the accounts of special science naturalness—especially since the discussion of them was so sparse. But, putting this aside, there are reasons why people might be doubtful of the general strategy. In this section I'll consider a few such objections.

(1) Objection: is this approach to answering the question ad hoc? Perhaps we can point to accounts of naturalness which explain the existence of regularities, but are these accounts well motivated? Or are they just cooked up in order to resolve this puzzle?

Response: as I just noted, there were two parts to the strategy in the paper. The first part, perhaps the more important one, is transforming the question about the existence of the special sciences into a question about naturalness. This transformation is not ad hoc. But what about the second part of the strategy—are the relevant accounts of special science naturalness unmotivated? When we describe the move in the bluntest way possible—saying that you should give an account of natural properties which builds in there being regularities about those properties—then it can seem ad hoc. But such accounts of special science naturalness are independently very attractive. As we noted, on both the PDA and the explanatory clustering approach properties count as natural if they play certain important roles in our scientific theorizing. Once we deny primitivism about special science naturalness then an account which looks to features of our scientific theorizing becomes very attractive.

What's more, there is something slightly odd about the accusation that such accounts of naturalness might be cooked up to resolve this puzzle, because it's not clear that's a bad thing. This existence of high-level regularities in a physical world is a deep, substantial puzzle, and if an account of special science naturalness can help resolve this puzzle then that's some motivation to accept the account.

(2) Objection: but aren't such accounts of special science naturalness very strange? In particular, they imply that intuitively unnatural properties count as natural in some possible worlds, because in those possible worlds such properties happen to be part of an explanatory cluster/be part of simple and informative ways of describing the world.

Response: the objector here is correct—on these approaches it is contingent which properties are natural. On the PDA, which properties are natural depends upon what the best way of summarizing the facts about the world is. This is clearly contingent. Similarly, on the explanatory clustering account which properties are natural depends upon how facts about those properties are explanatorily connected. Again, this is clearly contingent. And it likely will be the case that properties which are very unnatural in our world will count as natural on other worlds. In this way these accounts of special science naturalness differ substantially from the traditional primitivist approaches to naturalness which make facts about naturalness necessary.

But this isn't, I think, a bad result. Notice that many special science properties can seem unnatural when first introduced, and only come to seem natural once we are familiar with the theories in which they are embedded. For example, considered in itself, *gene* might seem fairly unnatural. Notice, for example, that the set of things that might realize *gene* is extremely disparate. It is only by understanding the role that *gene* plays in our theorizing—by seeing how genes are related to traits organisms possess and to inheritance of those traits—that we come to find the concept natural. That an intuitively unnatural property can come to seem natural at a world when we see how it is useful for theorizing at that world is, I think, not surprising.

(3) Objection: let's assume that the argument does successfully establish that we shouldn't be surprised that there are high-level regularities about relatively natural properties—even if the world was very different, there would still be high-level regularities about natural properties. However, the argument implies that if the world were very different, the natural properties would be very different.

This argument, therefore, leaves something important unexplained: why are there regularities about the properties that are interesting or salient to us? In our world there are lots of high-level regularities, and this needs explaining. But also, lots of those high-level regularities are about properties that seem relatively interesting or salient to us. That hasn't been explained—since in most worlds there will be regularities about properties that are natural in that world, but those properties will be very strange and uninteresting to us.

Response: I'm going to give a three-fold response.

Firstly, the properties that we find interesting or salient are not independent of my account. Part of why we find certain properties interesting is precisely because of the features that make them count as natural on the PDA or the explanatory clustering accounts—the way that they play an important and useful role in our scientific theorizing. Again, *gene* is not something that we are antecedently interested in—it only becomes interesting and salient because of its role in explaining and summarizing what we see. So there is a common explainer of its naturalness and us being interested in it.

But surely this isn't all there is for a property to be interesting or salient to us—surely some properties are interesting prior to our theorizing. Why are there regularities about such pre-theoretically interesting properties? Well, it's not at all clear that there are many regularities about such properties. For example, *love* is particularly interesting to us, and not in a way that depends upon our scientific theorizing, but we don't seem to have stable regularities about it. Similarly, I think, for other pre-theoretically interesting high-level properties.

(4) Objection: doesn't this make the existence of higher-level regularities too easy? This approach implies that pretty much whatever the world is like there will be higher-level regularities about relatively natural properties.

Response: this is an interesting objection. Earlier I noted that it's a problem if we give an explanation that makes the existence of special science regularities seem too fragile and surprising. Is it also a problem if we give an explanation that makes the existence of special science regularities too robust and unsurprising? Perhaps it is.

I certainly feel the force of this thought when we are considering the question of why the world is, at the fundamental level, regular. If someone attempted to explain why the fundamental level of the world is regular in a way that implies that there would be such regularity pretty much whatever the world is like, then I would be inclined to reject this explanation. I would be inclined to think that such an explanation can't really be explaining the right thing—whatever it's explaining, it's not the intuitive notion of the world being regular at the fundamental level.

So I understand if someone feels the same with respect to the question of why there are high-level regularities. But here is a potential difference. It's common to make the realist assumption that at the fundamental level there are some basic, metaphysically privileged properties. Properties out of which everything else is built. And when there are such metaphysically privileged properties the key question is why there are regularities about those properties. The type of strategy developed in this paper clearly cannot explain why there are regularities about some pre-identified set of metaphysically privileged properties.

However, it's also common to think that things are different at higher-levels—there are not higher-level properties that are pre-selected with some metaphysical "glow." In fact, our starting assumption of a strong kind of physicalism seems to rule this out. The specialness of certain higher-level properties must flow from the physical makeup of the world.

If this is right then with respect to the higher level the key question is not whether there are regularities about some pre-selected metaphysically privileged properties, but about why there are any such regularities of the type that we see in the special sciences. And this question is, I think, more amenable to the type of explanation given here.

Of course, people might disagree with this thought about the metaphysical difference between fundamental and high-level properties, even though it does fit with our initial assumption of physicalism. Clearly this is not the place to litigate those issues. But if you think that there are no metaphysically privileged fundamental properties then we may be able to explain why there are fundamental-level regularities in a way very similar to the strategy developed here.[9]

7 Conclusion

Why there are special science regularities is a deep puzzle. One strategy is to explain this in a bottom-up, piecemeal way—showing how particular regularities follow from the physics. But to explain the regularities we have is not to satisfyingly explain why there are regularities at all.

I've suggested, then, that there is a place for a more "metaphysical" style of explanation. And I outlined how that might go. The question of why there are special science regularities is, I've argued, intimately related to the question of what special science naturalness is. We can, therefore, attack the question of why there are such regularities via considering what naturalness is. Certain reductive views of naturalness plausibly make it unsurprising that there are regularities about the natural properties.

References

Albert, David Z. 2000. *Time and Chance*. Harvard University Press.

Batterman, R. W. 2000. "Multiple Realizability and Universality." *British Journal for the Philosophy of Science* 51 (1): 115–145.

Bhogal, Harjit. ms. "Special Science Naturalness." Available at: https://harjitbhogal.com/Special%20Science%20Naturalness%20Website.pdf.

Bhogal, Harjit. 2023. "The Package Deal Account of Naturalness." In *Humean Laws for Human Agents*, edited by Christian Loew, Siegfried Jaag, and Michael Townsen Hicks, 145–167. Oxford University Press.

Bhogal, Harjit. 2020a. "Coincidences and the Grain of Explanation." *Philosophy and Phenomenological Research* 100 (3): 677–694.

Bhogal, Harjit. 2020b. "Difference-Making and Deterministic Chance." *Philosophical Studies* 178: 2215–2235.

Block, Ned. 1997. "Anti-Reductionism Slaps Back." *Philosophical Perspectives. A Supplement to Noûs* 11: 107–132.

Dilthey, Wilhelm. 1894. *Ideen über Eine Beschreibende Und Zergliedernde Psychologie*. Verlag der Königlichen Akademie der Wissenschaften.

Dorr, C., and J. Hawthorne. 2013. "Naturalness." *Oxford Studies in Metaphysics* 8: 3–77.

Fodor, J. A. 1974. "Special Sciences (or: The Disunity of Science as a Working Hypothesis)." *Synthese* 28 (2): 97–115.

Fodor, Jerry. 1997. "Special Sciences: Still Autonomous After All These Years." *Philosophical Perspectives. A Supplement to Noûs* 11 (January): 149–163.

Garfinkel, Ala. 1981. *Forms of Explanation*. Yale University Press New Haven.

Hitchcock, Christopher, and James Woodward. 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Noûs* 37 (2): 181–199.

[9] In fact, via personal correspondence I gather that Loewer is attracted to this view.

Jackson, Frank, and Philip Pettit. 1992. "In Defense of Explanatory Ecumenicalism." *Economics and Philosophy* 8 (1): 1–21.

Kim, Jaegwon. 1992. "Multiple Realization and the Metaphysics of Reduction." *Philosophy and Phenomenological Research* 52 (1): 1–26.

Kitcher, Philip. 2001. *Science, Truth, and Democracy*. Oxford University Press.

Lewis, David. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343–377.

Lewis, David. 1986. *On the Plurality of Worlds*. London: Basil Blackwell.

Loewer, Barry. Forthcoming. *Fire in the Equations*. Oxford University Press.

Loewer, Barry. 1996. "Humean Supervenience." *Philosophical Topics* 24 (1): 101–127.

Loewer, Barry. 2001. "Determinism and Chance." *Studies in History and Philosophy of Science. Part B. Studies in History and Philosophy of Modern Physics* 32 (4): 609–620.

Loewer, Barry. 2007. "Laws and Natural Properties." *Philosophical Topics* 35 (1/2): 313–328.

Loewer, Barry. 2009. "Why Is There Anything Except Physics?" *Synthese* 170 (2): 217–233.

Loewer, Barry. 2020. "The Package Deal Account of Laws and Properties (PDA)." *Synthese* 199: 1065–1089.

Schrenk, Markus. 2006. "A Theory for Special Science Laws." In *Selected Papers Contributed to the Sections of GAP.6*, edited by H. Bohse and S. Walter, 121–131. mentis.

Sider, Theodore. 2011. *Writing the Book of the World*. Oxford University Press.

Strevens, Michael. 2008. *Depth: An Account of Scientific Explanation*. Harvard University Press.

Weslake, Brad. 2010. "Explanatory Depth*." *Philosophy of Science*, Ezproxy.library.nyu.edu, 77 (2): 273–294.

Wilson, Robert A. 1994. "Causal Depth, Theoretical Appropriateness, and Individualism in Psychology." *Philosophy of Science* 61 (1): 55–75.

Windelband, Wilhelm. 1904. *Geschichte Und Naturwissenschaft*. Humboldt-Universität zu Berlin.

Woodward, Jim. 2001. "Law and Explanation in Biology: Invariance Is the Kind of Stability That Matters." *Philosophy of Science* 68 (1): 1–20.

Yablo, Stephen. 1992. "Mental Causation." *The Philosophical Review* 101 (2): 245–280.

17

Why High-Level Explanations Exist

Michael Strevens

1 High-Level Explanation and Semi-Detachment

A simple evolutionary explanation invoking natural selection might go by way of a model structured as follows. The model contains variables representing the numbers of two variants in the relevant population, the fitness of each of these variants, and some assumptions about reproduction, including the way in which the variants are passed on from parent to offspring—in the simplest case, by an asexual mechanism in which the offspring has the same trait as its parent. Run the model, and the variant with the higher fitness comes to dominate the population.

More complex and interesting evolutionary models have more structure and more complications. Reproduction might be sexual, with the traits in question depending in perhaps elaborate ways on the genome. There might be many, or even an infinite number, of traits, if a phenotype can take on any of a range of characters (colors, shapes, etc.). The fitness of a trait might depend on the frequency of other traits in the population, or on other environmental variables subject to change in the course of the evolutionary process. Nevertheless, these more complex models have much in common with the simple model and with a vast array of other models used in high-level explanations.

On the one hand, they incorporate very precise representations of the systems whose dynamics they purport to capture. Evolutionary models may turn on tiny differences in fitness that slowly, but inexorably, work themselves out in favor of the fitter trait.

On the other hand, the precision in these representations falls well short of capturing the true causal intricacy of the system. Changes in a biological population are an aggregate of individual births and deaths, and these depend on such minute matters as the relative positions of predator and prey at particular times, which are in no way represented by the corresponding models. The models are in an important sense precise—they represent small differences, and what they predict and explain can turn on the details of those differences—but the precision is an entirely *high-level* precision. It manifests at the level of fitness, a parameter capturing a general fact about a phenotype in an environment, not a particular fact about the

Michael Strevens, *Why High-Level Explanations Exist* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Michael Strevens 2024. DOI: 10.1093/oso/9780192862945.003.0018

course of some specific organism's life as it makes its way around the environment. When such a model is put to explanatory use, the result is a paradigmatic high-level explanation.

Explanations having this high-level character proliferate in the sciences. An economic model of inflation may represent overall inflationary expectations, but not the expectations of individual economic actors, let alone their myriad other beliefs and goals. A model in population ecology, like many evolutionary models, may represent birth rates and death rates but not the facts that determine which organisms reproduce and when they die. A model of a chemically oscillating system, such as the Belousov-Zhabotinsky reaction, may represent the concentration of chemicals at each point in a shallow solution, yet leave out the motions and interactions of individual molecules which constitute the reaction. Even the most advanced weather forecasting models exclude a tremendous amount of causally relevant information. Many of these models manage passably accurate predictions all the same.

This chapter poses the question of how it is possible to have satisfactory explanatory models that operate mostly or wholly at the high level, omitting much of the causal detail that propels the systems in question along the trajectories that they take through the space of possibilities—omitting, that is, the eatings and matings of individual organisms, the decisions of individual buyers and sellers, the movements of individual molecules.

The answer to this question has two parts. The first part more carefully characterizes the independence or irrelevance of lower-level detail that makes high-level modeling feasible—a kind of autonomy that I call *semi-detachment* of the high level from the lower level. My characterization will not be especially formal or exacting, and indeed it does not depart in any especially notable way from many other thinkers' attempts to describe a certain kind of autonomy a process may have from the minutiae of its causal implementation. (I am thinking here of the work of Garfinkel (1981), and more recently Batterman (2002, 2021), Woodward (2021), and Robertson (forthcoming), along with the remarks of scientists such as Simon (1996), the luminaries of the Santa Fe Institute (Cowan et al. 1994), and Goldenfeld and Kadanoff (1999)). Certainly, the differences between my characterization and that of, for example, Woodward or Robertson will be of little importance in the execution of the second and rather more substantive part of this paper.

That second part sets out to provide an explanation of why semi-detachment is so widespread—an explanation that is all the more valuable because certain prima facie considerations suggest that it should be very rare indeed. Both for its intrinsic interest and because the practice of causal modeling in many disciplines hinges upon semi-detachment, this is a topic of immense significance. I would even go so far as to say that it is of comparable importance to, and has much greater practical significance than, philosophy's grand old problem of induction. Yet on the whole it has received only fleeting attention from philosophers of science.[1] I hope to change that.

2 Semi-Detachment Characterized

A high-level behavior of a system is *semi-detached* from its lower-level foundations if the behavior can be predicted accurately using only properties of the system that are themselves high level. Additionally—since I am interested in prediction that is also explanatory—I require for semi-detachment that the predictive high-level properties are *difference-makers* for the behavior in question, and that the prediction proceeds by deriving the behavior as a causal consequence of these difference-makers. (I put aside, then, "models" that make use of high-dimensional statistical correlations detected by machine learning to make their predictions.) The remainder of this section will offer remarks on this definition, along with glosses of important terms such as "level," "behavior," and "accuracy."

Let me begin with a couple of important observations about the notion of semi-detachment as a whole. First, semi-detachment is a property of a system's behavior, not of the system itself: one high-level behavior of a system might be semi-detached, another not. Second, semi-detachment is a kind of autonomy, makes possible a certain kind of multiple realizability, and has a kinship to notions of emergence such as that suggested by Wilson (2010). I name it using my own term of art to set it off from the many other senses of autonomy and emergence to be found in the philosophical literature. As observed in the previous section, it is nevertheless not something original to my own thinking; Garfinkel is an important precursor and Robertson and Woodward are fellow travelers. Third, there is inevitably a certain degree of vagueness and gradedness to the notion of semi-detachment. For simplicity's sake, however, I will tend to talk as though it is either present or not.

On to the notion of "levels." I don't assume any particular philosophical account, and certainly no particular metaphysics, standing behind talk of levels. We agree, I take it, that the population of a certain species in an ecosystem is a higher-level property of the system than the spatial location of an individual member of the species. That sort of agreement is sufficient for what I want to say in this chapter.

[1] I must make an exception for the case of universality in certain physical systems, memorably discussed by Batterman (referenced above) and his many commentators, and for the vast literature on the foundations of statistical mechanics (surveyed magisterially by Sklar (1993)). But the question of the general prevalence of semi-detachment—not just in physics, but in the various sub-domains of biology, in psychology, and in the social sciences—has been raised and pursued by only a handful of philosophers, such as Strevens (2003, 2005), Loewer (2009), and Bhogal (this volume). Among general-audience writers, Cohen and Stewart (1994) and Gribbin (2005) have also emphasized the importance of the question.

Definitions aside, a few important remarks about levels. First, “high” and “low” are relative in my usage. The position of an individual organism is a low-level property in the context of modeling in population ecology, but not low at all in the context of statistical physics.

Second, as should be clear from the foregoing, when I talk about a high-level model, I mean a model in which not only the target behavior—the behavior that is supposed to be modeled—but also all other information about a system represented by the model concerns high-level properties. Such properties include, most significantly, statistical or aggregate properties of the system’s parts: the number of organisms in a certain population or sub-population in an ecosystem, the average kinetic energy per degree of freedom of the molecules in a chemical solution, the mean income of wage-earners in a certain age bracket and socioeconomic group, and so on. Other high-level properties are what you might think of as high-level background conditions: the ambient temperature, or hours of sunlight in the day, or central-bank interest rates.

Third, information about high-level properties is always at the same time about low-level properties, because the high-level properties of a system are in some sense—a sense that may vary with the property—determined by the low-level properties. It is the disposition of individual gas molecules that determines the pressure of a gas, for example, and the existence of individual members of a species that determines that species’ population. To say that a model contains only information about high-level properties, then, is not to say that it contains *no* information about low-level properties, but rather that it contains only as much information about low-level properties as is entailed by the information it provides about high-level properties.

Fourth, throughout this chapter I assume that the high-level behavior of a system is ultimately determined by its low-level properties. Changes in the population of an ecosystem are brought about by, or depend on, the interaction of individual organisms: matings, eatings, and so on. The diffusion of one gas through another is brought about by the movements and collisions of individual gas molecules. And so on. For the purposes of this chapter, there is no need to understand this as a consequence of some entirely general reductionist principle (though I will admit that I am partial to such). It is quite enough that in the systems that I am most concerned with in what follows—especially biological populations and gases—it is true. In any case, the interest of the notion of semi-detachment, and thus the scope of the chapter, is limited to systems in which this low-level determination of high-level behavior is found. That is no great restriction, however, as such systems form the bulk of the subject matter of the special sciences.

What, next, is a “behavior”? Let me take my cue from science’s model-builders: it is the sort of change (or lack of change) in a system that modelers build their models to predict or explain. To characterize behavior, then, we should look at modelers’ predictive or explanatory goals. I don’t intend to pursue that project very far, but I do want to emphasize an important consequence of this way of thinking about behaviors: modelers do not normally aspire to model high-level behavior (or perhaps any behavior) with exactitude. To put it another way, what they aspire to model is not exact behavior but approximate behavior.[2]

The inexactness can take several forms (here focusing for simplicity's sake on deterministic models). First, even when a model, on the face of things, traces an exact behavior, modelers do not take that behavior at face value. For the right gases in the right situations, it is appropriate to use the ideal gas model to predict or explain, say, the change in pressure that is caused by a certain change in temperature. The ideal gas model is exact: there is no limit to the precision it offers, if given precise input. But no one takes this precision seriously. Because the modeled gas is not ideal, we expect its behavior to deviate a little from the model's ostensible prediction, and such deviations are not considered predictive or explanatory failures. The model is treated as a success when its predictions are approximately correct.

Second, in many applications, modelers do not expect models to be even approximately correct on every occasion. The predictions of the ideal gas law, or, for example, Fick's laws of diffusion, may in principle deviate profoundly from a gas's actual behavior (though the probability of this happening in any particular instance is vanishingly low). The laws are nevertheless regarded as excellent models of the systems in question.

In short, a predictively and explanatorily successful model of a system will typically capture that system's behavior only approximately, and may occasionally miss wildly.[3]

These remarks make it possible to characterize semi-detachment in somewhat more exacting terms. Like any attempt at philosophical precision, this one will not be free of artificiality and idealization, but I trust that it will bring a degree of clarity that justifies the cost. Let me suppose that predictive and explanatory models of high-level behavior consist of representations of possible states of the relevant system along with generalizations about how these states change over time. A simple model in population ecology, for example, might represent population levels, coefficients of reproduction and predation, a habitat's "carrying capacity" for each population, and a set of equations relating these properties so as to characterize the way that the populations will change over time. Let me suppose further that the generalizations are deterministic and exact, so that for

[2] For an overview of scientific modeling from a philosophical perspective, see, for example, Morgan and Morrison (1999), Weisberg (2013), and Frigg and Hartmann (2020).

[3] Because such models are not predictively perfect, it might be supposed that they are not explanatorily perfect, either. I do not rule out this possibility—that certain models that achieve greater accuracy by bringing in low-level detail are more explanatory than the high-level models that are my concern in this paper—but nor do I endorse it.

any given specification of the states (the parameters and variables), the generalizations will issue an exact representation of the state of the system at any subsequent time.

Such a model, then, is a deductive system. The question of what it has to say about a system's behavior is a matter of interpreting the exact representations it makes of the system's state over time. For the reasons given above, a modeler will tend not to take such representations literally. If the model says that the population of rabbits in one year's time will be 300, the modeler might understand the prediction as follows: *very likely*, the rabbit population in one year's time will be *approximately* 300. The modeler's goals will dictate the extent of these tolerances, and therefore the circumstances under which they regard the model as satisfactory.

When the model performs well given the relevant tolerances, I say that it is predictively and explanatorily accurate. Accuracy, then, does not require exactly 300 rabbits; it requires approximately 300 rabbits, and the occasional complete miss is typically allowed. (Of course, we would not say that the model is accurate on that occasion; what I mean is that such mishaps are not inconsistent with saying that the model is accurate in general.) My proposed use of the term "accurate" is, I should perhaps add, merely an expository convenience.

To return to the definition presented at the beginning of this section, a high-level behavior of a system is semi-detached just in case there is a purely high-level description of the system that captures enough information about the behavior's difference-makers to model the behavior accurately.

The changes over time in the population of a certain ecosystem, for example, are semi-detached from the low level if it is possible to build a model, incorporating representations of only high-level properties such as population number, that accurately models those changes. Putting accurate initial conditions into such a model must, then, consistently result in a prediction that conforms at least approximately to the actual subsequent behavior of the system, in most instances.

Semi-detachment is quite independent of scientists' aims and beliefs. Though it is because of our modeling practices, and in particular because of our extensive reliance on black-boxing and our other uses of abstraction and idealization, that we have so great a need for semi-detachment, the existence of semi-detachment is not determined by our practices but rather by objective, worldly properties of the modeled systems themselves. Were these systems not to exhibit semi-detachment, a model that omitted certain low-level details would thereby omit details that make a critical difference to the behavior of the target system, and so would neither reliably predict nor fully explain that behavior.

Semi-detachment is therefore a great boon to science, indeed, an essential condition for the kind of high-level predictive and explanatory modeling without which the special sciences could not exist.

3 The Puzzle of Semi-Detachment

Semi-detachment, if not ubiquitous, is certainly not uncommon—so I have insinuated. That fact may seem as mysterious as it is convenient. Low-level details habitually have a high-level impact: the relative position of a certain fox and rabbit is exactly the sort of thing to make a difference to the overall population of foxes and rabbits. Indeed, changes in population are determined by nothing but individual births and deaths, and therefore, it would seem, by the sort of fine-grained detail that high-level models of population ecology pass over in silence.

How, then, do things work out so happily? Why are our efforts at high-level modeling so often successful, both explanatorily and predictively? How does all the low-level detail, much of which causally contributes to high-level goings-on, conspire to cancel itself out, to add up, in all its causal potency and fecundity, to nothing? Or rather, nothing above and beyond its aggregate, its statistical manifestation at the higher level in terms of population numbers, average kinetic energy per degree of freedom, mean income, and so on?

A preliminary step in answering this question is to observe that the conception of a model's accuracy at the core of semi-detachment allows for the occasional gross deviation. If it is true that a flap of the butterfly's wings, or a twitch of the rabbit's ears, can cause a system to veer far from the trajectory that it would otherwise be expected to follow, then perhaps we can accommodate such paroxysms, if infrequent, under this escape clause. In effect, we are allowing that the high-level behavior we seek to capture with our models is an indeterministic behavior, a regularity marred by the occasional glitch. Such liberality makes it easier to find a suitable model.

But a certain tolerance on the modeler's part, though important, does not go to the heart of the matter. Exceptions aside, the modeler in their quest for accuracy demands high-level regularity, which is to say, behavior that can on the whole, if only approximately, be derived from high-level information alone. In order for there to be that kind of high-level regularity, the low-level aspects of a system's makeup that go unmentioned in a high-level model must either make no contribution whatsoever to a system's high-level behavior, or a contribution that is so consistent, so uncomplicated, that its net effect can be determined using high-level information alone.

The first option is not realistic, given that low-level details such as the relative position of individual predators and prey or gas molecules decide the difference between death and survival, collision and unimpeded travel, and these events in turn—deaths and collisions—are what drive changes in high-level properties such as an ecosystem's population or a gas's concentration. So it must be the second: all of that causally pertinent low-level complication and chaos somehow sums to a dependably rather simple ebb and flow at the high level.

The world need not have been so cooperative. Consider John Conway's Game of Life, a simple set of rules for cellular automata capable of generating a multifarious assortment of patterned and unpatterned behavior. Changing the state of a single cell in Conway's Game can send the system on a trajectory utterly different at even the highest level of description from the trajectory it would have traced without the change (Figure 17.1).

Why is real life so much simpler than the Game of Life? Why does the pattern of population change in a habitat of rabbits and foxes depend only on a few high-level variables, rather than varying with the starting position of this rabbit or that fox? Why, in natural selection, does one trait that differs only slightly from another reliably outcompete the other, rather than the race's being decided by the spatial orientation of some particular organism or other when the trait first appears?

One simple and straightforward explanation is that we are cherry-picking. Even in the Game of Life, there are many regularities to be found. Gliders, for example, move in a predictable direction at a predictable speed, provided that they do not encounter any other life. High-level models of the Game of Life may not be possible in general, but they are possible in certain carefully circumscribed circumstances.

Might our special sciences be selective and opportunistic in the same way? Might high-level modelers with a nose for semi-detachment converge on the few places where it is found, giving the impression that the property is widespread when in fact it is rare but well attended? Might the roving spotlights of scientific research pick out semi-detached high-level behaviors not because, wherever you point the light, you will find semi-detachment, but because the lights are manipulated so as to pick out nothing else?

Scientists will ever, of course, incline toward low-hanging fruit. A close look at high-level modeling practices suggests, however, that such bounty is remarkably common. Consider, for example, the way that students are taught to apply

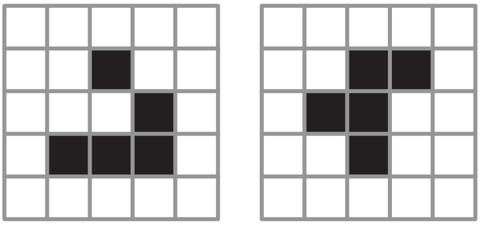

Figure 17.1 In Conway's Game of Life, the figure on the left, a "glider," keeps moving to the bottom right, cruising off to the edges of the universe (unless it encounters other life along the way). The figure on the right, the R-pentomino, if left to its own devices takes 1,103 generations to settle down to a stable state. By that time it has sent off six gliders and is composed of 116 cells. Similar figures behave quite differently, and have quite different end states—or, like a glider, never settle down at all.

the models of statistical physics, not tentatively and selectively, but enthusiastically and indiscriminately. Statistical models are simply expected to work in any of the enormous range of systems to which they ostensibly apply—and on the whole, they do.

Perhaps even more striking, because of the structural complexity and diversity of the systems involved, is the widespread applicability of high-level modeling strategies in biology. The models of population ecology are deployed wherever there are populations to model—whether to represent foxes preying on rabbits or malaria parasites on human children. The models of population genetics, similarly, are put into action wherever evolution is going on.

It cannot be said that these models invariably succeed as predictors the first time around. But when they fail, the modeler's assumption of semi-detachment is virtually never questioned. Instead, it is assumed that the model in question does not contain enough high-level information: ecological models may therefore be enhanced by building in representations of sub-populations, such as different age groups, or population genetics models by building in representations of more complex types of genetic interaction or mating preference. Modelers facing difficulties, in other words, double down on semi-detachment, proceeding as though models that contain sufficient high-level precision and detail will in the end attain some measure of predictive adequacy. This is, if anything, the reverse of cherry-picking: the modeler is not scanning the world for a few choice opportunities to apply their favored technique, but rather choosing to apply their technique anywhere and everywhere with an almost guileless confidence that is, in fact, rewarded over and over.

In short, while it is true that a great number of physical and biological systems have yet to be tested, those to which high-level modelers have applied their craft have tended, on the whole, to yield to the high-level approach—not because modelers are scrupulously selecting their targets, but because their habitual expectation of semi-detachment is satisfied far more often than not.

4 Explaining Semi-Detachment

4.1 Canceling Out and the Statistical Approach

Let me introduce the canceling-out explanation of semi-detachment by looking at simple monatomic gases. The molecules in a gas confined in a box careen around in the most chaotic way, sending each other flying in every direction. The overall state of the gas consists in nothing more than the positions and velocities of these molecules. Yet the aggregate of this intensely disordered motion—the high-level behavior of the gas—is a paradigm of order. A gas released into a box will almost immediately settle into an equilibrium state in which its molecules are approximately evenly distributed through the available space, and the speeds of the molecules conform approximately to the Maxwell-Boltzmann distribution (Figure 17.2). Shift a molecule to the right as you release the gas and (almost certainly) it will make no difference: the gas will head to the same equilibrium state and stay there for as long as you care to watch.

The reason that low-level chaos adds up to high-level simplicity is revealed by kinetic theory: the chaos (with very high probability) cancels out. A small change in the position of a molecule can, and usually will, result in that and many other molecules' later having completely different positions and velocities from those that they would have otherwise had. But from the high-level perspective, such changes are mere fluctuations, which in the aggregate are overwhelmingly likely to more or less balance each other in a way that results in a steady mean whose value is quite independent of the low-level details.

It is possible to understand the basic mathematical principle at work without plumbing the depths of statistical mechanics. Think of a coin tossed many times in succession. After a certain initial period, the frequency of heads will settle down to an equilibrium state in which it fluctuates, very slightly, around one half. The outcomes of the tosses are about as disordered as they could be, but that very disorder means that they balance out so as to ensure (with just the slightest chance of sizable deviation) a stable high-level behavior.

In a gas, the changes in molecular position and velocity that arise from motion and collision cancel each other out in the same sort of way. For every molecule that makes its way toward the left-hand side of the box, there is very likely another that makes its way toward the right-hand side, and so on, so that the overall distribution of molecules throughout the box remains the same. Likewise, for every molecule that, thanks to a collision, undergoes a sudden increase in speed, there will very likely be one that undergoes a sudden decrease in speed so as to replace it in the overall speed distribution shown in Figure 17.2. (The Maxwell-Boltzmann distribution is the equilibrium distribution precisely because it uniquely equalizes the rates of outflow from and inflow to each point along the speed axis.)

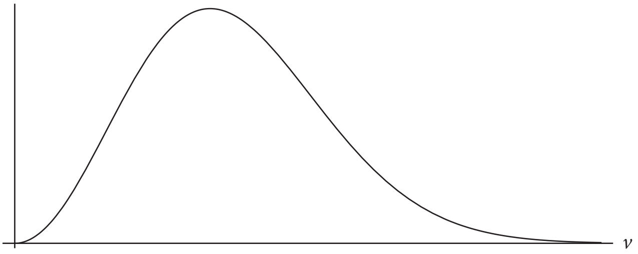

Figure 17.2 The Maxwell-Boltzmann distribution, showing the distribution of speed (i.e., velocity magnitude) for a gas at equilibrium.

In short, a promising approach to explaining semi-detachment invokes the canceling out of otherwise causally relevant low-level details, in the way described by statistical models of coin tossing and the kinetic theory of gases.

That observation prompts two questions. First, to what extent can we explain semi-detachment in other kinds of systems, such as ecosystems, using this approach? Second, even where technically feasible, are these explanations any good? They may work mathematically, but do they capture the real reasons for semi-detachment?

The answer to the first question is an unequivocal yes. I will argue by example. Consider a simple Lotka-Volterra model of a predator/prey ecosystem, representing the way in which interacting prey and predator populations—foxes and rabbits, say—change in time.[4] The model contains just two variables, representing the populations in question—*x* for the number of prey and *y* for the number of predators. It is encapsulated in two differential equations:

$$\mathrm{d}x/\mathrm{d}t = Rx - Pxy$$

$$\mathrm{d}y/\mathrm{d}t = Qxy - Sy$$

The first equation states how the rate of change of the population of prey ($\mathrm{d}x/\mathrm{d}t$) depends on the current population of prey and predators, as well as two constants *R* and *P*. The equation's first term *Rx* represents the rate of increase in the prey population due to reproduction; the constant *R*, then, is proportional to the mean number of offspring produced per member of the prey population. The second term represents the rate of decrease in the prey population due to predation. That number is proportional to the population of both predators and prey: the more prey milling about, the more are taken by any individual predator, and the more predators on the prowl, the more the total number of prey taken—or so the model asserts.

The second equation states how the rate of change of the population of predators depends on the current population of predators and prey. The first term is the rate of increase in the predator population due to reproduction, conceived of as proportional to the available food, and therefore to the rate of predation—the flip side of the second term in the first equation. The second term represents the rate of decrease in the predator population due to death from any causes. The constant *S* therefore captures the predator death rate per capita.

Such elementary models typically cannot predict with much accuracy changes in the modeled populations, but they can predict and explain qualitative features of population change—for example, the tendency of many ecosystems to manifest

[4] For a philosophically inflected discussion of this model and its uses, see Weisberg and Reisman (2008).

a robust equilibrium, in which prey and predator populations are roughly constant, or the tendency of a general biocide (a deleterious change in the environment that equally kills predators and prey) to result, in the medium term, in a larger relative proportion of prey. With respect to certain qualitative high-level behaviors, then, the Lotka-Volterra model is accurate in my sense. (More sophisticated high-level models, I should add, can do a much better job of quantitative prediction, and are used as forecasting tools by park administrators, fisheries managers, and so on.)

The accuracy of Lotka-Volterra models illustrates the semi-detachment of many high-level behaviors in a simple ecosystem. The qualitative effect of a general biocide, for example, depends only on the aspects of the modeled system that are explicitly represented in the model. It does not depend, then, on low-level details such as the positions of particular organisms (or, indeed, on many high-level aspects of the system).

How to understand this lack of dependence, this semi-detachment? The mathematics of the Lotka-Volterra model itself is deterministic through and through. Nevertheless, we can explain semi-detachment as a matter of low-level causes "canceling out" by conceiving of the underpinnings of the behavior described by the model statistically, that is, along the same lines as we conceive of coin-tossing or the statistical mechanical underpinnings of the physics of gases.

Consider, for example, the predator death term in the second equation, namely, Sy. It is rather natural to interpret the constant S as representing the probability that any particular predator dies over the course of a time interval of length 1 (in the units in which the model measures time). You might think of it this way: as a predator makes its way through the system, it has a certain chance of contracting a fatal infection or suffering a fatal accident. It's a random matter, rather like drawing balls from an urn. Pick the black ball, and it's game over.

Drawings from an urn are like coin tosses. Over time, black balls will be drawn with a frequency that represents their proportion in the urn. To predict whether or not the next ball to be drawn will be black, we need a compendious description of the state of the urn and fearsome amounts of computation. But to predict the frequency with which black balls will be drawn, we need only know the proportion of balls that are black. The remainder of the facts about the state of the urn impact the frequency only as fluctuations that, in the medium term, with high probability cancel out.

Likewise for the ecosystem: though to predict the death of any particular organism we might need to know an immense amount about its position, condition, and the positions and conditions of every other organism in the system, to predict the overall rate of death we need only a relatively small amount of high-level information, determining the overall degree of danger. All of the hurly-burly of everyday life can, then, be compressed into a single number analogous to the urn's black ball proportion, a high-level property of the system as whole that accurately models the death rate. Everything else affects the death rate only as a series of fluctuations that very likely, if the population is not too small, balance out.

Or consider the term in the first equation that represents the rate at which prey are eaten, namely, *Pxy*. We can think of the vicissitudes of predation in the following way. Predators roam the habitat, looking for prey. The more prey there are, the more likely a predator is to encounter one and to eat it. It is as though each predator is drawing balls from an urn, determining whether or not they find a prey in some particular patch of the habitat. The urn contains a ball for each patch, either fleshy pink for "prey" or white for "empty." The more prey there are in the system, the higher the proportion of pink balls. Each predator, then, will have a chance of catching a prey proportional to the number of prey. Now, this sampling is going on for every predator. The expected number of prey caught in a given time interval, then, is proportional to the number of predators (the number of animals sampling the urn) and the number of prey (the number of pink balls in the urn, given that the total number of balls—representing approximate positions in the habitat—is fixed).

Over time, and if the populations are large enough, the actual rate at which prey are caught will very likely closely track the expected number. Thus it will be equal to a constant multiplied by the predator and prey populations—or *Pxy*, as in the model. The low-level complexity inherent in the biological dynamics of predation contributes to the predation rate only, again, in the form of fluctuations, much like the fluctuations in the frequency of "pink" caused by individual urn-drawings, canceling out and leaving behind a rate of death by predation that is dictated by high-level properties of the system alone.

The same kind of story might be told for any high-level model of this sort—any high-level model consisting of difference or differential equations representing change as a function of high-level quantities, when that change is driven by the outcomes of numerous low-level events such as predatory pounces or molecular collisions. We have a general strategy, then, for understanding the semi-detachment of high-level behavior—not, perhaps, applicable to every model in the special sciences, but apt for a vast range, including models of natural selection, economic equilibrium, statistical physics, and more.

4.2 Are Canceling-Out Explanations Valid?

The strategy is only as good, however, as its assumptions. Are the foragings and ultimate fates of predators like urn drawings or coin tosses in the relevant way? There are some strong prima facie reasons to think not, and thus to doubt the canceling-out explanation of semi-detachment, answering my second question above—is the explanation any good?—in the negative.

An initial concern is that the statistical strategy ascribes objective probabilities where there are none. Arguably, the low-level dynamics of predator-prey interactions are nearly deterministic, in the sense that any fluctuations bubbling up from the quantum level have little or no effect, characteristically, on their outcomes. And where there is determinism, some would say, there can be no physical probability or "chance" (Schaffer 2007).

But this concern should not delay us. The low-level dynamics of a coin toss are known to be nearly deterministic, yet a statistical model provides a satisfying explanation of coin-tossing's tendency to deliver a stable frequency in the medium to long term, showing how low-level detail cancels out. This may not be physical probability or "chance" in a grand metaphysical sense, but our goal is not metaphysical enlightenment but physical understanding, and statistical modeling is a fine way to understand canceling out and stable frequencies in deterministic systems, as in the coin toss and as in my interpretation of the underpinnings of the Lotka-Volterra model.[5]

Some more telling objections to the canceling-out explanation raise concerns not about the strategy of statistical modeling as such, but rather about the particular statistical assumptions that go into the story. I understood the predator death term *Sy* in the Lotka-Volterra model as a consequence of a fixed probability that any predator in the system would die over the course of the relevant interval of time (one unit of the time variable *t*). But surely it is not the case that every predator has an equal chance of dying in a given span of time. Very young, very old, or very sick organisms are more likely to go under.

This is all true. A better understanding of the predator death term is as follows. Assume that the age and health profiles of the population are in equilibrium. The proportion of organisms that are very young, then, remains the same even as the absolute number of organisms changes, as does the proportion of organisms that are old or sick. Then we can understand the constant *S* in the death term not as a single death probability valid for every organism, but rather as a weighted mean of such probabilities—in technical terms, the marginal probability of predator death. Given the assumption of age and health equilibrium, the marginal probability will not change (unless the underlying probabilities change), and so we can safely put it to work to model the predator death rate.

What if, in a given ecosystem, age and health are not in equilibrium? Then the simple model examined above is not valid. What we need instead is a model that tracks the age and health structure of the population, with separate variables representing the number of young, adult, and old predators, sick and healthy predators, and so on for any factor that affects the probability of death, along with coefficients representing the differing probabilities. And population ecologists do

[5] As argued by Strevens (2003), Myrvold (2021), and others, following a tradition initiated by von Kries (1886) and Poincaré (1896).

indeed employ such models when necessary. Observe that like the original model, they represent the system in wholly high-level terms: the number of young predators is as much a high-level property of the population as its total number. The behavior of the system characterized in these somewhat finer-grained terms, then, is a high-level behavior, and one that floats free of further low-level detail—one that is semi-detached.

A second concern is that the probability of death for a particular organism depends on low-level details that, by contrast with age and health, cannot be packaged into high-level statistics in the way just proposed. Some predators are killed by falling trees. You might suppose, then, that an organism's chance of death will depend on its proximity to dead or dying trees. As it gets closer to such a tree, the probability of death inches up (*ceteris paribus*). Likewise, the probability of a prey's death surely depends on its proximity, at any moment, to a hungry predator. In short, death probabilities depend on the details of relative positions at particular times, information that is never represented, even statistically, in high-level population models.

This variation in the probability of death might be handled using something similar to the equilibrium posit above—an assumption, that is, that each prey spends about the same proportion of its time near and far from hungry predators, and that each predator spends about the same proportion of its time near and far from weak-rooted trees. That seems a promising strategy. It is, after all, on the whole a matter of pure chance whether an organism finds itself in this sort of peril: there is no reason why one organism should receive any greater exposure to danger than any other (putting aside phenotypical differences that could be captured by a population structure model).

A more rigorous (if far from deductive) argument to this effect is developed in Strevens (2003), chapter four. There I develop the idea that very small, short-term fluctuations in a creature's day-to-day meanderings will, thanks to various kinds of sensitivity to initial conditions, have a randomizing effect on matters such as proximity to danger. It really is as though organisms of a given sub-population in a given habitat are all making drawings from the same urn, subject to the same probability of drawing the red ball of peril.

Surprisingly, then, it turns out that the chaotic aspect of life—the sensitive dependence of important outcomes on small, seemingly insignificant matters of fact—which was portrayed as a threat to semi-detachment in Section 3, is an important factor in ensuring the canceling out that makes semi-detachment possible. It is not the only factor, or else the Game of Life would yield to the high-level approach, but it is an essential part of the story.

A final concern with the statistical thinking behind the canceling-out approach is that it assumes stochastic (i.e., statistical) independence where there might seem to be none. Return for a moment to the case of coin tossing. Each toss in a series is stochastically independent of the outcomes of the other tosses, meaning that the other outcomes make no difference to the probability of heads on that particular toss: it is one half whether preceded by a head, a tail, or some longer and more elaborate sequence.

Stochastic independence is a part of what explains the tendency for outcomes on a series of coin tosses to cancel out, resulting in a frequency for heads that in almost every case fluctuates just very slightly around one half. If the canceling-out explanation of semi-detachment is to be applied to the behavior represented by the Lotka-Volterra model, then, a corresponding independence assumption is required. The probability that one predator dies should be independent of the death of another; the probability that one prey is eaten should be independent of whether any other prey suffers the same ending.

Such an assumption may appear dubious, for two reasons. First, in the case of the tossed coin and other such randomizing devices, the rationale usually given for supposing stochastic independence is causal independence: the probability of heads on a given toss does not depend on the outcomes of other tosses because it is causally disconnected from the others. That certainly does not hold in the ecosystem: prey are continually interacting with each other and with predators, and vice versa.

Second, there are positive reasons to think that certain of these causal dependencies undermine stochastic independence. A predator can eat only so much at a time. If one prey is consumed, then, the probability that another will be consumed shortly afterward surely decreases, with one less hungry animal on the prowl.

In fact, however, an assumption of *approximate* stochastic independence—that is, outcomes' having very little if not zero impact on other probabilities—can be sustained in the ecosystem. It is true that in many cases a predator's feeding will make all the other prey a little safer. But in an ecosystem of any size (and we need the size in any case to get canceling out), the effect is small. The predation terms in the Lotka-Volterra model remain roughly correct.[6]

That accounts for one possible source of dependence. But the interaction between the animals in an ecosystem is so pervasive, so intense, and so potentially consequential—I remind you again that very small variations in position can make the difference between life and death—that it is impossible to maintain that the creatures persist in a state of even approximate causal independence. Their causal connections are thick and tangled; there is virtually no causal independence whatsoever. (Here sensitivity to initial conditions resumes its former role as a spoiler.)

Causal independence may be (at least in normal circumstances) sufficient for stochastic independence, but in other work I have shown that it is far

[6] To put it another way, it is as though the model supposes that the notional urn-drawing is with replacement (a ball is put back in the urn immediately after it is drawn), when in fact it is not. If an urn contains sufficiently many balls of each color, however, the difference between sampling with and without replacement is on average quite small.

from necessary. Consider two tossed coins that collide in midair. They interact substantially—certainly enough to make a difference to whether the coins land heads or tails. Yet under a wide range of conditions, the outcomes are stochastically independent. Learning how one coin lands gives you no help in predicting how the other lands (Strevens 2015). Further, a strong case can be made that the relevant probabilities in an ecosystem work in much the same way. (Strevens (2005) gives an accessible visual treatment; Strevens (2003) goes deeper.) The canceling-out explanation is therefore viable after all.

4.3 Expanding the Circle

I've said quite a bit about population change in ecosystems, but what about other behaviors and other kinds of system? Economic systems? Chemical systems? Social systems of various stripes?

The canceling-out approach is applicable—generalizing from the cases discussed above—if the following conditions hold. First, the semi-detached high-level behavior in question involves high-level properties whose dynamics are determined by the outcomes of many individual low-level events (births, deaths, changes in molecular velocity, coin tosses). Second, these low-level outcomes have probabilities whose values depend only on high-level properties (or fixed features) of the system in question, as the probability of prey death depends only on the overall number of predators and prey (and perhaps a fixed distribution of age and health). Third, the low-level outcomes are at least approximately stochastically independent.

That is a sufficient condition; perhaps something along the lines of a canceling-out explanation can be given in related cases as well. Or perhaps in some systems the canceling-out approach will constitute one part of a multipronged strategy. For social systems, for example, canceling out might explain certain regularities in the social background which, in tandem with the plasticity of human behavior, account for further high-level regularities in the course of human affairs.

In other cases, the best explanation of semi-detachment may have nothing to do with canceling out. If we want to understand semi-detachment across the board, then—if we want to understand the viability of high-level modeling and the power of high-level explanation wherever they are found in the special sciences—we will need a veritable toolkit of techniques. Canceling out, however, will surely constitute one of this toolkit's principal explanatory instruments.

References

Batterman, R. W. (2002). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction, and Emergence*. Oxford University Press, Oxford.

Batterman, R. W. (2021). *A Middle Way: A Non-Fundamental Approach to Many-Body Physics*. Oxford University Press, Oxford.

Cohen, J., and I. Stewart. (1994). *The Collapse of Chaos*. Viking, New York.

Cowan, G., D. Pines, and D. Meltzer (eds.). (1994). *Complexity: Metaphors, Models, and Reality*. Addison-Wesley, Reading, MA.

Frigg, R., and S. Hartmann. (2020). Models in science. In E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Spring 2020 edition. <https://plato.stanford.edu/archives/spr2020/entries/models-science/>.

Garfinkel, A. (1981). *Forms of Explanation*. Yale University Press, New Haven, CT.

Goldenfeld, N., and L. P. Kadanoff. (1999). Simple lessons from complexity. *Science* 87: 87–89.

Gribbin, J. (2005). *Deep Simplicity: Bringing Order to Chaos and Complexity*. Random House, New York.

von Kries, J. (1886). *Die Principien der Wahrscheinlichkeitsrechnung*. Mohr, Freiburg.

Loewer, B. (2009). Why is there anything except physics? *Synthese* 170: 217–233.

Morgan, M. S., and M. Morrison (eds.). (1999). *Models as Mediators: Perspectives on Natural and Social Science*. Cambridge University Press, Cambridge.

Myrvold, W. (2021). *Beyond Chance and Credence: A Theory of Hybrid Probabilities*. Oxford University Press, Oxford.

Poincaré, H. (1896). *Calcul des Probabilités*. First edition. Gauthier-Villars, Paris.

Robertson, K. (Forthcoming). Autonomy generalised; Or, why doesn't physics matter more? *Ergo*.

Schaffer, J. (2007). Deterministic chance? *British Journal for the Philosophy of Science* 58: 113–140.

Simon, H. A. (1996). *The Sciences of the Artificial*. Third edition. MIT Press, Cambridge, MA.

Sklar, L. (1993). *Physics and Chance*. Cambridge University Press, Cambridge.

Strevens, M. (2003). *Bigger than Chaos: Understanding Complexity through Probability*. Harvard University Press, Cambridge, MA.

Strevens, M. (2005). How are the sciences of complex systems possible? *Philosophy of Science* 72: 531–556.

Strevens, M. (2015). Stochastic independence and causal connection. *Erkenntnis* 80: 605–627.

Weisberg, M. (2013). *Simulation and Similarity: Using Models to Understand the World*. Oxford University Press, Oxford.

Weisberg, M., and K. Reisman. (2008). The robust Volterra principle. *Philosophy of Science* 75: 106–131.

Wilson, J. (2010). Non-reductive physicalism and degrees of freedom. *British Journal for the Philosophy of Science* 61: 279–311.

Woodward, J. (2021). Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance. *Synthese* 198: 237–265.

18

A Democracy of Laws

Michael Townsen Hicks

1 Introduction

The distinct sciences are characterized by a combination of independence and mutual constraint. Different sciences employ different methodologies: they develop a conceptual structure, use that structure to formulate explanations through laws and explanatory models, and hold those concepts and laws accountable to the world through experimentation and observation. The generalizations they arrive at support counterfactuals and feature in explanations.[1] Despite this independence, the sciences exercise mutual constraint on one another: even counterfactual disagreements between sciences show that at least one set of laws contains a falsehood. And these sciences exhibit a hierarchical explanatory structure: inter-scientific explanations flow up, from more to less fundamental sciences. Accounting for these four features is the job of a philosophical account of law in the special sciences. This problem has generally been approached as the problem of *reduction*: which sciences *reduce* to which? Specifically, do all sciences reduce to physics? And how is the relationship of reduction to be understood?

Understanding the problem as a problem of reduction is mistaken for two reasons. Firstly, it biases the discussion against views which emphasize the methodological independence of the sciences. Secondly, it creates the illusion that we are looking for a simple yes-or-no answer. Disagreements over whether, e.g., mere supervenience is sufficient for reduction distracts us from the underlying features of the relationship between sciences that need to be explained. To avoid these

[1] Woodward (2003), Ch. 6, denies that the generalizations of the special sciences are laws; in doing so, he rejects the notion that only laws are counterfactually invariant, and that only laws are available for use in explanations. Woodward's reasons are simple: according to standard accounts of law, laws must be exceptionless. But the generalizations which feature in special scientific explanation are not spacio-temporally unrestricted and have exceptions. For an argument that there are no laws in biology, see Beatty (1995); for a response, see Mitchell (2000). Like Woodward, I have no truck with a verbal dispute about the word 'law'. Here, and throughout this paper, I will use 'law' to refer to those counterfactually robust generalizations that can underwrite predictions and feature in explanations. Claiming that the generalizations of the special sciences are not laws will not remove the burden of explaining these features of those generalizations, and so will not (by itself) solve the coordination problem. Consequently, I will not address the question of whether laws must be spacio-temporally unrestricted.

Michael Townsen Hicks, *A Democracy of Laws* In: *Levels of Explanation*. Edited by: Katie Robertson and Alastair Wilson, Oxford University Press. © Michael Townsen Hicks 2024. DOI: 10.1093/oso/9780192862945.003.0019

confusions, I'll call the puzzle posed by the relationship between the sciences the *scientific coordination problem*: how are various scientific disciplines coordinated with one another?

In this paper, I'll present a new solution to the problem of scientific coordination. But first, in Section 2 I'll identify two strains among extant solutions to the problem of scientific coordination. Following Weslake (2014) I call these the *imperialist* and the *anarchist* solutions to the scientific coordination problem. The imperialist sees the special sciences as a consequence of fundamental physics; the laws of the special sciences are laws because they can be *derived from* or *grounded* in the laws of physics. This strong reductionist view seeks to make every explanation an explanation from physics. The anarchist, on the other hand, denies that the sciences are so tightly connected. Rather, she sees them as each unifying a body of facts, or cataloguing the dispositions of properties. Both of these views fail to solve the scientific coordination problem; the imperialist fails to account for the independence of the sciences, and the anarchist fails to account for their mutual, asymmetric dependence.

I'll then (Section 3) offer a third view, which I call *the democratic view*: on my view, the various sciences work together to generate a set of laws, which together maximize a certain sort of informativeness. Rather than being a Better Best System view, this is a Big Best System view (in the terminology of Callender and Cohen (2009)): scientific laws earn inclusion in the Big Book of Science by adding sufficiently to its informativeness without overly complicating the overall picture. Like the Better Best System, this view is a version of the Humean Best Systems Account (BSA) of laws. On the BSA, the laws of all science together form a simple, informative set of axioms useful for limited physical agents. Though my view is similar to other Humean views of special scientific laws (such as Callender and Cohen (2009, 2010) or Schrenk (2014)), it differs in important respects. First the informativeness of this mega-lawbook is evaluated holistically—the laws of all sciences are taken to form a coherent informative system, rather than being evaluated independently with respect to their domains. However, because various scientific disciplines are epistemically isolated in a way in which I will make more precise in Section 3.2, they add to this lawbook semi-autonomously. The view I advocate has the central advantages of both the imperialist and the anarchist views. Like the anarchist, and unlike the imperialist, I hold that the laws of the special sciences are *made laws* in the same way that the laws of fundamental science are. Like the imperialist, but not the anarchist, I hold that the laws of physics are fundamental, and that there is an asymmetry between the special sciences and fundamental physics.

I then (Section 3.3) round out the case for democracy by showing how it deals with various features of the relationship between fundamental and special scientific laws. First, it gives us a new way of understanding *ceteris parabis* conditions. Second, it provides an understanding of the 'naturalness' of special scientific vocabulary that does not rest on its definition in more fundamental terms. Third, it allows the Humean to make sense of the idea, championed by Woodward (2013), that special science laws have different degrees of counterfactual robustness.

Though the divide between imperialism and anarchism largely crosscuts views about the metaphysics of laws, the proposal I offer depends on features of the Mill-Ramsey-Lewis regularity theory of laws. This is a Humean view: it relies on no fundamental notions of necessity or dependence. But non-Humeans will find much here to like: Humeans take the epistemic role of laws to be *constitutive* of natural lawhood. That is, they believe that laws support counterfactuals and provide explanations because of their epistemic utility, not vice versa. A modalist about laws, who takes laws to have either irreducible nomic or metaphysical necessity, will still need to understand the epistemic utility of laws, and so can tack this account on to their more metaphysically[2] robust account as an explication of the epistemology of laws. And many modalists about law take the only truly necessary laws to be those of fundamental physics; such a metaphysician of law can accept this view as an account of the laws of the special sciences while denying that it is a sufficient account of lawhood simpliciter.

2 The Imperialist and the Anarchist

In what follows, I will first sketch out two families of views concerning special scientific lawhood and then show how individual philosophers fit into one or another camp. It's worth noting that views about the relationship between physics and the special sciences largely crosscut views about the metaphysics of laws; although ultimately I favor a broadly Humean view of laws, my criticisms of the current theoretical space of possibilities do not rest on any metaphysical scruples.

To help illustrate the difference between the anarchist and the imperialist, and later to elucidate the democratic view, I'll make use of an idealized epistemic agent. She needs, unlike us, to have a vast capacity for absorbing and combining information from various sciences. But we will not assume that she is logically omniscient, or that she, like Laplace's demon, is able know everything about the state of the world (though she might), nor will we assume that inference is for her without computational costs. Some of these details of our agent will be fixed by the various purported solutions to the scientific coordination problem. We can refer to her

[2] Here I'm lumping together a variety of non-Humean views; these include Armstrongian (Armstrong 1983) necessary connections, which are a higher-order, intrinsically modal relationship, Birdian (Bird 2007) essential natures of properties, and sui generis modal facts like those advocated by Maudlin (2007) and more recently Chen and Goldstein (2022) and Adlam (2022). These views have something in common: they rely on a novel bit of ontology which has an intrinsic modal character to capture the necessity of laws of nature.

as a FISA: a Fairly Ideal Scientific Agent.[3] In these ways, she is a less-than-ideal Bayesian agent, and her credences encode the laws of various sciences.

If one of our laws says that if A then B, FISA's credence in B conditional on A ($F(B \mid A)$) will be 1. But the laws FISA responds to need not be deterministic: if our laws are statistical, this will be reflected in her credences. So if it is a law that agents who are asked to memorize a ten-digit number are more likely to utter racial slurs than those who have no number to remember, her credence $F(slur \mid \neg number)$ will be less than her credence $F(slur \mid number)$.[4]

I will evaluate imperialism, anarchism, and democracy with respect to four features of the relationship between physics and the special sciences (briefly introduced in Section 1). These desiderata must be a bit vague; different views about the relationship between the sciences should be allowed to provide slightly different accounts of what, for example, the asymmetric dependence between physics and biology amounts to.

- METHODOLOGICAL INDEPENDENCE: each science is able to formulate generalizations and support them evidentially via induction, and each science is able to determine its own conceptual structure. So, for example, Carnot, Clausius, and Kelvin were able to discover the laws of thermodynamics well before its concepts were located in a more fundamental physical theory by Boltzmann.
- COUNTERFACTUAL ROBUSTNESS: the generalizations of the special sciences are counterfactually robust: that is, they both support counterfactuals and hold in a variety of counterfactual situations—including, plausibly, counterfactual situations in which the laws of lower-level sciences do not hold. So, for example, the laws of supply and demand both support counterfactuals in this world and would hold even in worlds with very different physics (provided they had societies engaging in economic activities), and the Hardy-Weinberg law, which tells us that, in certain conditions, genotypes remain constant across generations, would hold[5] even if genes were realized by something other than DNA.
- MUTUAL CONSTRAINT: distinct sciences cannot make inconsistent predictions, including predictions about what would occur in merely counterfactual situations, and cannot provide inconsistent constraints on belief or credence (as Meacham (2014) argues—though see Hoefer (2014) for a response).

[3] The strategy of explicating views of laws via an idealized scientist is becoming more common, and appears in Callender and Cohen (2010) and Hall (2015).

[4] Typically, discussions of objective probability assume that the objective probabilities are precise in situations in which they are defined. But this is not obviously the case for some special science generalizations: plausibly, some laws in the special sciences provide comparative relations between conditional probabilities without nailing those probabilities down. While I think that a complete account of special scientific law should be compatible with this (and believe that the view defended here is), addressing this issue is beyond the scope of this paper.

[5] And maybe does!

Closely related sciences are such that the entities studied in one science can be located amongst the entities studied in another, often via a functional reduction, or due to the asymptotic behaviour of the entities of the more fundamental science (as Batterman (2001) argues).

- ASYMMETRY: metaphysical or grounding explanations between sciences go in one direction only; this direction of explanation creates a hierarchy roughly lining up with the direction of mereological dependence, where the entities of higher-level sciences are made up of the entities of lower-level sciences. One way in which this asymmetry manifests itself is as follows: entities and behaviour at the higher level can be located amongst the entities studied at the lower level; higher-level regularities are often targets for explanation at the lower level. An excellent example of this is the reduction of chemistry to physics, where chemical kinds—elements—are taken to be arrangements of physical kinds—protons and neutrons. The stability of some arrangements of protons and neutrons but not others explains the limited number of elements; the physical properties of these arrangements, such as the allowable energy levels of electrons orbiting them, explain the chemical properties of the elements in question, such as electronegativity.

We are looking for a view of laws that explains these four aspects of the relationship between the sciences while retaining descriptive adequacy: the closer the laws posited by the view resemble those of our current sciences, the better. It should be believable that the solution under discussion is a view about the laws of our sciences—if the view does not allow some special scientific generalization to be a law, or requires us to add to the fundamental laws, this is a demerit of the view.

This is a defeasible requirement. For the laws we have now are not the final laws; and the divisions we now carve between our sciences are somewhat arbitrary. So a philosopher has it within her rights to argue that our final theory will have features no current theory has; and she may likewise argue that some laws which are currently considered to be in science A actually belong in science B—or that the division between science A and science B isn't the division we should be worried about. But as these conjectures about a yet undreamt final science move us further from the theories we have, her view loses more plausibility.

2.1 The Imperialist

The imperialist view holds that the lawhood of the laws of the special sciences derives from the lawhood of the fundamental laws[6] (or the laws together with 'robust'

[6] It's important to bear in mind here that we are discussing the dependence of *laws* on *laws*. Any physicalist philosopher is committed to some dependence of higher-level facts, including the facts about which generalizations are laws, on the physical *facts*. But this dependence need not go directly

initial conditions). The imperialist may hold that they can be *derived* from the fundamental laws, but she need not: she may hold instead that they are metaphysically necessitated by the fundamental laws, or that they are *grounded* in the fundamental laws—where A grounds B only if A metaphysical necessitates B and A explains B.

A prototypical—if dated—imperialist is F. P. Ramsey (1929), who held that there are three grades of law: fundamental laws, laws that are derived from the fundamental laws alone, and laws which are derived from the fundamental laws and some 'robust' initial conditions. We might add a fourth category, not available to Ramsey: laws derived from the fundamental laws and *a posteriori* necessities, like 'water = H_2O'.[7] Finally, we should remember that it is open to imperialists to add to the set of fundamental laws so that they have sufficiently strong implications for the special sciences.

Our FISA, according to the imperialist, starts with a set of fundamental laws. These laws may be sentences which together maximize strength and simplicity, as the Humean holds (Lewis (1980, 1983), Beebee (2000, 2006, 2011), Loewer (2007, 2008, 2009)), they may be generalizations which are backed by a relationship of necessitation between universals (Armstrong (1983, 1997)), or they may be sentences which describe the dispositional essences of the properties which feature in them (Ellis (2001), Bird (2007)). She then works out the consequences of these laws. On the most austere view, her conditional credences now encode the fundamental laws and the laws of the special sciences. I'll call this view the 'austere imperialist' view.

But on more permissive views, she isn't done. On what I'll call a 'permissive imperialist' view, all she has now are the fundamental laws. She may still either conditionalize on some special set of the initial conditions, or she may conditionalize on *a posteriori* necessities—typically property identities. Once she has done this, says the permissive imperialist, she has at her disposal both the fundamental laws and those of the special sciences.

It's worth noting here that the laws of the special sciences need not receive probability 1. Indeed, likely they should not. For the laws of the special sciences are not exceptionless, as are the laws of fundamental physics. So an adequate account of special scientific law, imperialist, anarchist, or democratic, ought to hold that the conditional credences assigned to the special scientific laws are less than maximal.

through the laws; the unifying claim of imperialism is that the *lawhood* of the special sciences is dependent directly on the laws of fundamental physics (together perhaps with some other physical facts).

[7] Of course there's a fifth possible type of law, one dependent on all three together: the fundamental laws, robust initial conditions, and *a posteriori* necessities. But these will not improve the situation for the imperialist: I will argue that neither initial conditions nor *a posteriori* necessities can ground the laws. If these cannot solve the scientific coordination problem on their own, neither can the two together.

Before we look at the problems with imperialism, we should note its advantages. Imperialism clearly and coherently explains two features of the scientific coordination problem: MUTUAL DEPENDENCE and ASYMMETRY. According to imperialism, the laws of the various scientific disciplines must be compatible because some of them are a consequence of others (together, for the permissive imperialist, with robust initial conditions or *a posteriori* necessities). If we discover a contradiction between the apparent predictions of two sciences, it's impossible that one of them is derived from the other. Consequently one of them must have the wrong laws.[8] And the ASYMMETRY of the sciences is neatly explained as well, because the laws of less fundamental sciences are a consequence of those of the more fundamental science, but not vice versa. The asymmetry of the sciences is just the asymmetry of deduction: the special scientific laws follow from those of physics, but not vice versa.

As to the METHODOLOGICAL INDEPENDENCE of the special sciences, the imperialist gets a weak pass. For the imperialist is not committed to our FISA actually representing scientific reasoning; we may not be able to perform the computations which FISA performs. She is fairly ideal, and so may be ideal in ways in which we are imperfect. So—perhaps—we with our limited cognitive resources are forced to engage in standard inductive reasoning to discover the laws of the special sciences, rather than simply deriving them from the laws of physics (together with whatever else). According to the imperialist, the fact that some special science generalization is inductively supported is strong evidence that it is a law, and so strong evidence that it is a consequence of the laws (and 'robust' facts) of physics.

This pass is a weak one. For the imperialist has given us no reason—at least not yet—to believe that the inductively supported generalizations of the special sciences will line up with those derivable from physics. Note that it is not enough for the imperialist to appeal to the counterfactual robustness of special scientific laws and claim that this robustness must come from the laws of physics. For the source of this counterfactual robustness is precisely what is at issue![9] Rather, she must provide some independent reason to believe that higher-level inductive reasoning will arrive at the consequences of physics, rather than some other generalizations.

Despite these successes, imperialism lacks the resources to explain COUNTERFACTUAL ROBUSTNESS while retaining descriptive adequacy. To see this, let's first examine *austere imperialism*. Austere imperialism holds that the laws of the special

[8] Imperialism doesn't hold that the mistaken science must always be the special science; we might take fundamental physics—such as quantum mechanics—and thermodynamics together to be fundamental, but see that the contradiction between physics + thermodynamics and geology, recognized by Kelvin in the nineteenth century, told against the then-dominant theory of physics rather than the then-dominant theory of geology. Because the geological laws yielded a different age for the earth than physics + thermodynamics, and because if B contradicts A it's not the case that A implies B (given that A is self-consistent), we know that one of A or B must *not* be a law. We don't know whether to take this as a modus ponens of $\neg$ B or a modus tollens of A.

[9] Loewer (2008) argues that, because the higher-level frequencies are determined by statistical mechanical probabilities, observations of higher-level frequencies give us evidence about the underlying fundamental probabilities. I will address this later.

sciences are a consequence of the laws of physics alone. We can see right away that austere imperialism will simply not do: for the laws of physics alone have too few direct consequences to underwrite all of the special science laws. And this is reflected in the structure of the laws of physics and the laws of the special sciences. The laws of physics are temporally symmetric, exceptionless, and deterministic.[10] The laws of the special sciences are temporally asymmetric, have exceptions, and are often statistical. So the special scientific laws could not be a result of the laws of physics on their own.[11,12]

Now consider the permissive imperialist who adds *a posteriori* necessities. It's not at all clear how this could help. For if the laws of physics are temporally symmetric, exceptionless, and deterministic, adding a metaphysically necessary lasso between these laws and some higher-level terms will not introduce an asymmetry, exception, or indeterminism. This, of course, is the source of all sorts of difficulties in the foundations of thermodynamics and statistical mechanics.

So to retain descriptive adequacy, the imperialist ought to become more permissive. She ought to include not only the laws of physics and *a posteriori* necessities, but also some 'robust' initial conditions. To make this work, she will need a clear notion of robustness: one which will lead to an explanation of the lawhood of special scientific laws. By adding facts about the past, and not the future, we can secure the temporal asymmetry, exceptions, and indeterminism of the special sciences (see Albert (2000)). But note that adding these initial conditions immediately makes this aspect of the scientific coordination problem more pressing: for if the initial conditions are not *themselves* laws, how can they make *other* generalizations laws? It seems that the imperialist must talk fast if she is to explain COUNTERFACTUAL ROBUSTNESS: it's clear how more fundamental laws can make special science laws necessary. But how can initial conditions make them necessary? After all, if the laws are deterministic, then the initial conditions together with the law imply everything; but not every fact is a special science law. The imperialist must explain pretty quickly what makes some features of the initial conditions special (it is just this issue that leads Beatty (1995) to argue that biology is without laws).

[10] Quantum mechanics, on either the orthodox or Ghirardi-Rimini-Weber (GRW) formulation, is indeterministic and temporally asymmetric. But this should not concern us: first, the orthodox interpretation is widely regarded to be inadequate, both in specificity (it posits collapses, but does not say when or how they occur) and in internal consistency (the indeterministic collapse postulate is in tension with the deterministic evolution of the wavefunction). Meanwhile, the GRW interpretation makes predictions which are distinct from those of orthodox quantum mechanics, but which enjoy limited empirical support. In either case, it's doubtful that the temporal asymmetry and indeterministic nature of quantum mechanics underlies the asymmetry and indeterminism in the special sciences. Finally, on either of the other two leading interpretations of quantum mechanics (Bohmianism and Everettianism), physics is deterministic and temporally symmetric.

[11] This is extremely clear if the dependence relation is something like derivability. But more permissive dependence relations, like supervenience or metaphysical grounding, face the same problem: if the supervenience base, or grounding facts, are temporally symmetric, exceptionless, and deterministic, how can they by themselves ground asymmetric indeterministic laws?

[12] For a more thorough and engaging discussion of this problem, see Loewer (2008).

The contention here is not that accidental facts never support counterfactuals. They do: the accidental fact that my favorite mug just appeared on a TV show makes it the case that *if I were to sell it on Ebay, I would make $70.* The worry is instead that the laws of the special sciences are robust in a way that these accidents are not. The fact that all of the coins in my pocket are quarters makes some counterfactuals true, but it's not the case that if this nickel were in my pocket, it would become a quarter. The laws of biology are not like this: it's true that if I were a bear, I would hibernate through the winter. This second class of counterfactuals, about what would occur under some manipulation, is the sort of counterfactual that can be grounded by laws but not accidents.[13]

Next, without a specification of which initial conditions are *robust*, the imperialist's solution to the problem of METHODOLOGICAL INDEPENDENCE is even more fraught. For whichever initial conditions she chooses, she will need to explain why *those* initial conditions, and not the others, make a generalization available for inductive discovery at the higher level. But there is no reason to believe that there is any set of conditions on robustness that will do this. In fact, there is reason to believe the opposite.

The challenge for an imperialist is to find a set of facts which (a) together with the fundamental laws ground the laws of the special science, (b) do not mistake accidents for laws at the special scientific level, and (c) are sufficiently counterfactually robust to underlie the counterfactual robustness of the special scientific laws. In order to satisfy the third desiderata, the facts appealed to by the imperialist must be in some sense unified. If they are not sufficiently unified, the imperialist view will lack the resources to explain the counterfactual robustness of these generalizations without succumbing to ad hockery. But these three conditions have yet to be met: imperialist theories either have too little in their grounding base or have too much. If they have too little, they can't explain the lawhood of all special science laws. If they have too much, they don't explain the accidental nature of special-scientific non-laws. Taking these two horns together, imperialists fail to identify a non-ad-hoc set of initial conditions to add to the base, and so they are unable to explain the counterfactual robustness of the special sciences in terms of the counterfactual robustness of physics. Consequently, they have failed to deliver on their imperialist promise.

2.1.1 Statistical Mechanical Imperialism

To see this, we may do well to examine one of the most worked out extant imperialist theories: that of Loewer (2008, 2009) (I follow Weslake (2014) in calling Loewer's

[13] The best way to characterize the counterfactual stability of the laws is disputable. Lange (2009) holds that the laws form a maximally subnomically counterfactually stable set—that is, they would hold under any counterfactual supposition which is not about lawhood and not inconsistent with them. Woodward (2013) has a slightly different, manipulationalist view about their counterfactual invariance conditions. I don't think I need to take a stand on which characterization is correct to make trouble for imperialists.

view ‘Statistical Mechanical Imperialism’). Loewer recognizes that initial conditions on their own cannot support counterfactuals; so he argues that some initial conditions ought to be included in the book of laws. Specifically, he thinks that, in addition to the laws of physics, our fundamental lawbook should include PROB, ‘a law that specifies a probability distribution (or density) over possible initial conditions that assigns a value 1 to PH [the initial low entropy condition] and is uniform over those microstates that realize PH’ (Loewer, 2008: 19). As this low-entropy initial condition is a law, it is just as able to underwrite counterfactuals as the other laws in our fundamental lawbook. And PROB, Loewer argues convincingly, deserves to be in our lawbook for the same reason other laws are: adding it dramatically increases the informativeness of the lawbook without unduly complicating it.

So far, Loewer looks to have solved the problems of austere imperialism without adding the paralysing complications of the permissive view. PROB is temporally asymmetric and probabilistic, and so can underwrite similar temporal asymmetries and probabilistic higher-level laws that don’t follow from physics alone. But because PROB (according to Loewer and Albert) is a law, it neatly explains the counterfactual robustness of its consequences.

Unfortunately, PROB and the laws of physics cannot save imperialism. They are, by themselves, too permissive: many generalizations will have high probability, according to them—more than are counted as laws by the special sciences. This is because many highly probable generalizations will be burdensomely gruesome: we can take any two special scientific laws, which we can assume are given a high probability by the Loewer-Albert system. We can then define gruesome predicates by pasting together terms from each law, and thereby arrive at a gruesome generalization at least as probable as the conjunction of the two laws. If they have a high enough probability, this generalization will also have a probability above whatever threshold we set for lawhood, but because of its gruesomeness, will not be a law (one option here would be to add higher-level natural properties à la Wilson (forthcoming)).

And there is no guarantee that the laws of the special sciences we have will be given a high initial probability by these two. To see this, consider a law of population genetics. Such a law will depend sensitively on contingent facts early in the evolution of modern animals (it is just this problem which is discussed in Beatty (1995)). But PROB does not give a high probability to these historical facts—or at least does not probabilify them over their alternatives. So if the story is that very high-probability facts are higher-order laws, this looks like it will predict the wrong laws historically; it is unable to distinguish the right laws as counterfactually robust as we had hoped.[14]

Loewer recognizes this, and the view he arrives at is closer to the permissive imperialist view: ‘The special science laws that hold at t are the macro regularities

[14] Of course, many of these facts currently have a high probability, since the universe updates its probabilities as events occur. But again, in this case they are not distinguished from non-lawful facts, which also are high probability now that they have occurred.

that are associated with high conditional probabilities given the macro state at t' (Loewer, 2008: 21). 'As the universe evolves … the probability distribution conditional on the macro state will also evolve.' We can illustrate this with our FISA as follows: she starts out with credence 1 in the laws of physics, and in the low-entropy macrocondition. Her conditional credences are uniform with respect to those microstates that realize the low-entropy macrocondition. As the universe evolves, our FISA conditionalizes on macroscopic information—that is, information about the positions of middle-sized dry goods, their temperatures and densities, locations and velocities. At any time, having conditionalized on all of the universe's macroinformation, those generalizations with high probability are the special scientific laws at that time.

Here we have a permissive imperialist view with a well-defined notion of robustness: the robust initial conditions are those which are encoded in the world's macrostate. But we can see immediately that this too is problematic: first, not all true macroscopic generalizations are laws; but all true macroscopic generalizations will get probability 1 on the scheme advocated by Loewer. Second, some true macroscopic generalizations will be laws despite *not* having high probability conditional on macroscopic information. Consider generalizations of population genetics. These are true because of some facts about the structure of the chemicals which convey our genes. But these chemicals are *not* macroscopic; they are microscopic. So they will not be conditionalized on by our FISA, and the generalization will not be a law.

Perhaps there is a way of tweaking the Albert/Loewer view to account for this; but I'm doubtful that there is an independently specifiable set of facts such that conditionalizing the uniform distribution over microstates on these facts will yield a high probability to all and only special scientific laws (a similar point is made forcefully by Frisch (2014)).

And this generalizes: for a permissive imperialist view to work, there must be some non-ad-hoc way of specifying which initial conditions are 'robust' enough to ground higher-level laws. Without such a specification, the imperialist has no way to distinguish laws from non-laws at the higher level. And without a way of distinguishing the laws from non-laws, we will not have the beginning of an explanation of COUNTERFACTUAL ROBUSTNESS and METHODOLOGICAL INDEPENDENCE. In order to explain why the special scientific laws are supported by induction and support counterfactuals, we must first distinguish between them and the non-laws, which are *not* supported by induction or counterfactually robust. The permissive imperialist cannot do this.

2.2 The Anarchist

The anarchist holds that the laws of the special sciences are laws for the same reason that the fundamental laws are. What makes the special science laws lawful? This question will be answered differently by different anarchists—Humean anarchists, like Craig Callender and Jonathan Cohen (2009, 2010) or Markus Schrenk (2008, 2014), claim that they provide the best systematization of facts in the language of their science (though, for Callender and Cohen, the choice of language is arbitrary or pragmatic). Anti-Humean anarchists, like Nancy Cartwright (1999), hold that the laws of the special sciences, like the laws of physics, encode dispositions or capacities which manifest in the controlled environments that that science studies. There are, according to Cartwright, no principles coordinating the laws outside of these controlled environments.

While Callender and Cohen and Cartwright agree that the laws *and* facts of the special sciences and physics depend on one another symmetrically if at all, this is not a requirement of anarchism. I will call anarchists of both law and fact 'radical anarchism'. Schrenk (2014) argues that the facts of special sciences supervene on those of physics, although the laws are independent; I will call this view 'moderate anarchism'.

According to the radical anarchist, our FISA will have a number of distinct, possibly incomplete credal functions available to her. Each of these will be defined over a different set of propositions: $F_{biology}(A \mid B)$, $F_{physics}(C \mid D) \ldots$. According to Callender and Cohen, A and B, C, and D are different propositions because they come from different ways of partitioning the space of worlds; there may be some overlap between, say, A and C, and there may even be a translation between the AB partition and the CD partition, but the probability functions are distinct and defined over different propositions. Which credal function FISA uses depends, according to Callender and Cohen, on which is easiest for FISA to apply to the situation at hand. Which evidence propositions are most easily verified in this situation? Which conditional probabilities are easiest to calculate?

Similarly for Cartwright. FISA will avail herself of a variety of disjointed credal functions, but instead of each being complete over a partition of the space of propositions, they will each be incomplete and only defined within certain controlled situations. So in situations in which $F_{physics}(A \mid B)$ is defined, $F_{biology}(A \mid B)$ is not. The situations in which physics yields a conditional probability are those with X-rays and scanning-tunnelling microscopes; the situations in which biology yields conditional probabilities are those in which groups of animals interact. Which credal function FISA uses will depend on the situation in which she finds—or creates for—herself.

It is compatible with anarchism that the *facts* at the special scientific level depend asymmetrically on the *facts* at the fundamental level; but anarchists deny that the *laws* so depend. Views of this latter sort—according to which the laws are in some way emergent, despite the dependence of the facts at the higher level on the facts of fundamental physics, are held by Fodor (1974), Lange (2009), Armstrong (1983), and Schrenk (2014). According to these philosophers, the independence of the higher-level laws arises because the laws of the special sciences describe patterns which are visible only at the coarse-grained higher level, are not the result of the laws of physics alone, are the result of the laws of physics together with any suitably special initial conditions, and are backed by modal facts (necessitation relations or irreducible counterfacts) which are independent of both the lower-level modal facts and the higher-level categorical facts. Because this version of anarchism allows some dependence between facts at different scientific levels, we will call it 'moderate anarchism'.

Both varieties of anarchism score well in accounting for the METHODOLOGICAL INDEPENDENCE and, at first brush, the COUNTERFACTUAL ROBUSTNESS of the generalizations of the special sciences. The counterfactual robustness of special scientific generalizations is explained in the same way as the lawhood of fundamental generalizations: either with sui generis modality or in terms of unificatory power. Similarly, the methodological independence of the special sciences is explained easily by the metaphysical independence of the laws. Special scientists are able to perform inductions in the same way physicists are because their laws are the same as those of physics.

Radical anarchism does poorly in accounting both for the MUTUAL CONSTRAINT and the ASYMMETRY of the special sciences and physics. On Cartwright's view, any two sciences don't attempt to describe the same world; rather, they make predictions about distinct controlled situations. No rules govern how they interact with one another, but plausibly the capacities of any science can overturn those of any other. So it's surprising that scientists seek information from one another, and that contradictory predictions are taken to indicate that one or another science's laws must be altered.

Radicals realize this; both Callender and Cohen and Cartwright argue that neither of these hold.[15] Unfortunately I do not have space to address their arguments here; so we will give them a demerit for failing to account for these relations, but note that this consequence of their view is not one these folks take to be a negative.

Moderate anarchism does better in explaining MUTUAL CONSTRAINT and ASYMMETRY of the scientific coordination problem. According to these views, constraint and asymmetric dependence arise from the metaphysical dependence of the *facts* of the special sciences on the facts of fundamental physics. We were understandably mistaken in our belief that these constraints held at the level of laws.

[15] Callender and Cohen reject asymmetry, but accept mutual constraint. On their view, each science forms a deductive system in an independent vocabulary. Because the vocabularies describe the same world, they must agree on the categorical facts of the world. Consequently, no generalization at any level can imply that another generalization is (actually) false. However, nothing in their view guarantees that the laws will agree on what happens in counterfactual situations: a systematization could rule that, for some merely possible event A, if A were to happen, then B would, while another could rule that if C were to happen, then D would, where A metaphysically entails C but B and D are mutually contradictory; if A does not occur, Callender and Cohen can't guarantee that this would not be the case. Similarly, they cannot guarantee that the chances assigned by various laws will yield compatible constraints on credence.

This view cannot be correct. For the laws of the special sciences have exceptions, and these exceptions can often be explained by the laws of lower-level sciences. In fact, in many (though not all) cases, the exceptions to a special scientific law can only be specified by appeal to a lower-level science. Whether or not the laws of the special sciences have *built in ceteris parabis* conditions,[16] specifying situations in which they do not hold requires us to take on board concepts *which are not a part of the special science in question.* The predictions of economics can be trusted *provided* an asteroid does not strike the market.

More subtly, explaining *for which* species the Hardy-Weinberg law holds can only be done by discussing properties of DNA; explaining which—highly unlikely—scenarios are entropy-increasing and so violate thermodynamics' second law can only be done only by citing the momenta of the particles underlying the system. But meteor impacts are not describable in the conceptual scheme of economics (we have astrophysics for that), describing DNA proteins requires chemical, and not merely biological, concepts, and discussing the (non-aggregate) features of the particles which make up a gas is outside of the conceptual sphere of classic thermodynamics (for a discussion of the differences between traditional *ceteris parabis* conditions and the sort of 'no odd realizers' condition which holds for thermodynamics, see Fenton-Glynn (2016)).

This observation allows us to recognize a problem for anarchism's explanations of asymmetry of COUNTERFACTUAL ROBUSTNESS. For though the anarchic view may be able to explain the force of the special scientific laws, it is unable to explain why their exceptions are often outside of the conceptual scope of the science in which they feature. The anarchist has no explanation of the fact that lower-level laws trump higher-level laws and can explain the exceptions to higher-level laws, but not vice versa.

All versions of anarchism face the *conspiracy problem* (see Callender and Cohen (2010) for a discussion). If the laws of physics and the laws of the special sciences are independent, how is it that they conspire together to produce a unified world? That is, why is it that the laws of physics somehow 'know' not to push elementary particles around in a way which violates the laws of the special sciences? And how do the special scientific laws, like those of psychology, fail to license violations of the laws of physics? The conspiracy problem is a challenge to the anarchist solution to MUTUAL CONSTRAINT; the anarchist claims that the sciences describe the same world, but if she is radical and holds that their laws are metaphysically independent, how do they combine to create a coherent world? This problem is even more pressing for chancy laws. Meacham (2014) shows that, on the sort of scheme

[16] There's a lively debate about this. For a discussion about whether *ceteris parabis* conditions are parts of the laws in which they feature, see, among others, Cartwright (1999), Mitchell (2000), Woodward (2003, 2013), and Hüttemann (2014)).

Callender and Cohen advocate, we will receive different advice on our probabilistic beliefs depending on which language we're using.

A distinct challenge for both anarchist views—but especially the moderate anarchist—lies in explaining the COUNTERFACTUAL ROBUSTNESS of the special sciences. We gave anarchists a strong pass on this earlier: their explanation of special scientific lawhood is, presumably, the same as their account of fundamental scientific lawhood. But just as the moderate anarchist cannot force the chances of the various sciences to align, they similarly cannot force the counterfactuals to align. For while the fact that these sciences all describe the same world requires them to agree about what *actually* happens, they may well disagree about what *would* happen if things were slightly different. This matters: we use laws to evaluate counterfactuals so that we can make decisions about what to do. These decisions in turn lead us to actualize some situations but not others. If laws disagree about what would happen if (for example) I pour more hot water into my mug, I will have no guidance about whether or not to do so. This fact may lead me to not pour the water; while the fact that I decide not to actualize this situation allows the laws to remain consistent, it also makes it the case that I missed out on some warmer tea.

Together, these problems create a dilemma for anarchist views. For the moderate anarchist: if the fundamental laws govern the fundamental facts, and the fundamental facts explain the special scientific facts, what is left for the laws of the special sciences to do? For the radical anarchist: if the laws all independently determine the facts, how do they manage to produce a consistent world? The more radical an anarchist is, the less she can explain MUTUAL CONSTRAINT. The more moderate she is, the less she can account for the COUNTERFACTUAL ROBUSTNESS of the special scientific laws.

The anarchist response is to claim that the special scientific laws *explain* in a way which is not reducible to the laws of physics. But note that this requires us to (a) take lawhood to be deeply tied to explanation, rather than governing, and (b) accept an overdetermination of explanation. While many philosophers, especially of the Humean strain, will not find either of these especially troubling, philosophers who take lawhood to be connected with governing, and who take explanation to be similarly tied to causation (rather than unification), may well reject these costs.

3 The Democratic View

I've argued that a successful solution to the scientific coordination problem cannot take the lawhood of the special sciences to be wholly dependent on the laws of the physics. And I've further held that the laws of each science cannot be made laws entirely by facts within the domain of that science. Both views leave one or more of our explananda—METHODOLOGICAL INDEPENDENCE, COUNTERFACTUAL

ROBUSTNESS, MUTUAL CONSTRAINT, and ASYMMETRY—unaccounted for. How, then, can these desiderata be met?

In this section, I'll present a view according to which the sciences work together to generate a unified body of knowledge. The generalizations in any science are laws, not because of their explanatory capacity given the facts of that science, or because of their relation to more fundamental generalizations, but because of their contribution to the informativeness of the total set of scientific laws. The mutual constraint the laws exercise on one another is a result of the fact that this informativeness is evaluated holistically: the laws of all sciences taken together contribute to the informativeness of our system. So they need to produce an internally consistent and mutually reinforcing set of predictions. And the independence of the various sciences is also accounted for: each science contributes laws to the overall system independently.

This view is a development of David Lewis's Best System Account of laws. In Section 3.1 I'll present this more holistic best system account. In Section 3.2 I'll show in more detail how special scientific laws add to the informativeness of a lawbook, and in what way this is mirrored in the credences of our FISA. Finally, in Section 3.3 and 3.4 I'll show how this view successfully accounts for some additional connections and differences between the special and fundamental sciences.

3.1 The Democratic Best System

Since the view I advocate builds on Lewis's Best System Account (Lewis (1983)) I'll briefly rehash his view here. The BSA holds that laws are the general truths of that axiomatic system which best combines simplicity, informativeness, and fit (where fit measures how closely chancy laws match the frequencies of the world). The motivation of the Best System Account is simple: we are interested in generalizations which can be used by us and give us a lot of information about the world. The virtues identified (strength, simplicity, and fit) are justified because they either measure how usable by us the set of generalizations are (this is what simplicity does) or because they measure how much information—either binary or probabilistic—the laws convey. However, as an impressive number of authors advocating Humeanism have noted, Lewis's understanding of strength, simplicity, and fit are in sore need of revision. I've already discussed how Callender and Cohen (2009, 2010) and Loewer (2007) depart from orthodox Lewisianism. But modifications are also suggested by Hall (2015), Hicks (2017), Hoefer (2007), Jaag and Loew (MS), and Woodward (2013). The upshots relevant for this view are twofold.

First, we are interested in finding the most informative lawbook we can. But we prefer that this information come in the form of widely applicable dynamical laws—laws which operate as functions from states of the world at one time to states of the world at another. Next, while scientists employ simplicity considerations in theory choice, they do so because simpler laws are better evidentially supported and lead to more accurate predictions.

Now, suppose we have a maximally informative fundamental dynamical law: one that tells us, given the precise fundamental state of the world at any time, exactly how that world will develop. How useful is that law? In how many situations will we be able to apply it? The obvious response is: surprisingly few. We will only be capable of employing such a law when we have precise information about our global surroundings. If we have only coarse-grained information about the boundary conditions of the system we're studying, such a law will provide us with precisely no useful predictions—there will be precise, fine-grained states of the system we're examining which develop in all sorts of bizarre ways, ways we cannot use the laws to rule out (this is the central argument of Albert (2000)).

The natural way to strengthen such a system, then, is to add more laws: dynamical principles which take *coarse-grained* information about the world and tell us what to infer about its *coarse-grained* future state. Of course, doing so will decrease the simplicity of the laws. And it may also decrease their accuracy: these laws of coarse-grained evolution may very well have exceptions, and when they do, we will be led astray. So each addition of coarse-grained laws has a cost to the overall utility of the lawbook. When will such additions be warranted? Plausibly, when the higher-level behaviour is sufficiently *novel* and *autonomous* (in the sense of Butterfield (2011)), such that it can't, or can't easily, be divined from the laws that are already on the books. And when the higher-level laws are sufficiently *accurate*: that is, using them to make inferences doesn't lead us astray (at least, not too much).

The key move here is to understand the informativeness of the laws as a bridge between the boundary information we bring to a situation and the predictions we can make using them. The laws are more informative if they lead us to more predictions, sure. But they are also more informative if they lead us across that gap from rougher ground.

The important balancing act that the democratic view maintains is this: though the laws of each science are added to the lawbook independently, on their merits as informative, accurate, and widely applicable generalizations, the book is scored holistically. This allows the sort of independence sought by Callender and Cohen's Better Best System account while avoiding the risk that the different sets of laws will conflict with one another. For if two laws which conflict are both added to the book, the laws as a whole will fail to make any predictions in any situations (or, equivalently, they will predict everything in every situation), and so their informational value to us will collapse.

3.2 Democracy in Detail

To see how this works with more precision, let's return to our FISA and consider her interests in formulating lawful generalizations. She is interested in discovering the most informative set of *conditional* probabilities $F(P \mid B)$ where P is a prediction and B is a set of boundary conditions. Her lawbook can be strong in two ways: first, it can be strong by being accurate: the conditional probabilities can be such that $F(P \mid B) \approxeq 1$ for situations in which both P and B hold, and $F(P \mid B) \approxeq 0$ when B and $\neg P$. But her lawbook can also be strong by being more applicable: that is, it can give her predictions for a wider range of situations, represented by the boundary conditions B. Call the first variety of strength *accuracy*, and the second *comprehensiveness*.

Accuracy and comprehensiveness trade off against one another: a lawbook can gain comprehensiveness by applying to situations with less uniform phenomena, although by doing so it will be unable to provide as accurate predictions of their behavior. Maximizing the combination of these virtues is hindered by the fact that the laws need to be in some sense *repeatable*: they must be formulated in such a way that multiple distinct situations have, according to the laws, the same boundary conditions, and so the laws must yield the same predictions in those situations. This is a requirement if the laws are to be discovered and evidentially supported by induction. The laws are generalizations which we can learn in one context and apply to another.

Thinking of strength in this way combines the notions of strength and fit: although accuracy is a rough analogue of fit, and comprehensiveness is a rough analogue of strength, neither precisely maps on to the Lewisian notion. Repeatability plays part of the role in trading off against comprehensiveness that simplicity does in the traditional best system, but it is not a perfect match; and for our laws accuracy trades off against comprehensiveness and repeatability together. We can have more accurate probabilities that are tailored to each experimental situation, but they will not be repeatable; we can have a probability function which is highly accurate but only by excluding some situations, and it will not be comprehensive; and we can have a repeatable, comprehensive probability function that moves further away from 1 for some true predictions and further away from 0 for some false ones.

Fundamental physics is extremely accurate. But it is not comprehensive. For any maximally fine-grained propositions B_0 and P, a deterministic physics will give $F_0(P \mid B_0) = 1$ if and only if B situations lead P and assign $F_0(P \mid B_0) = 0$ if not. But physics is silent about less fine-grained propositions: suppose B is the proposition that the temperature of a gas is T, its volume is V, and its pressure is p. Even given the identification of temperature, volume, and pressure with the kinetic properties of a collection of gas particles, our FISA will not have enough information to supply the boundary values of the variables in her fundamental dynamical law.

Our agent's information about the boundary conditions of a system need not be maximally fine-grained. But if she conditionalizes on the proposition that cream has been added to her coffee, what should she think will happen—will they mix? Unfortunately, nothing. For even adding *a posteriori* identities relating physical properties to thermodynamic properties, we still will not arrive at a prediction for the temperature of the gas: there are physical states compatible with the boundary conditions which are temperature increasing, and physical states compatible with the boundary conditions which are not (these latter states involve the energy of the particles which are impacting the boundary of the gas, and so increasing the pressure, are quite different from the energy of those particles in the interior, which contribute to the temperature). So while a set of laws in terms of maximally fine-grained propositions may be accurate, it will not be comprehensive.

To increase the comprehensiveness of the laws, we may add laws which take us from course-grained states—like temperature and pressure—to other course-grained states. This is the project of thermodynamics. Or we can add a probability function over the fine-grained states which is invariant under the fine-grained dynamics. This is the project of statistical mechanics. In either case, our predictions will diverge from perfect accuracy, so we will lose some accuracy in the overall system. But we will be able to apply the laws in many more situations. I claim that each scientific discipline increases the comprehensiveness of the overall lawbook. By adding more laws at a higher level, we increase the comprehensiveness of the overall system with some moderate sacrifice to its accuracy. The view here builds on the work of Handfield and Wilson (2013).[17]

Let's see how our FISA will behave on this way of understanding the laws. She will begin with a set of fundamental laws; she'll work out the consequences of these laws, and generate a probability function $F_0(P \mid B_0)$. This probability function will be incomplete; it will only be defined for maximally fine-grained propositions P and B, and while it will have precise conditional probabilities in those variables, it will be indeterminate concerning the unconditional probabilities of various possible initial conditions.[18] So FISA will see if there is a set of more coarse-grained variables in which she can formulate fairly accurate and repeatable laws. She'll add these to her lawbook, and work out the consequences, arriving at an extended credal function $F_1(P \mid B_1)$. But this credal function still won't be defined over propositions at all levels of grain, either because these coarse-grained laws don't imply *more* coarse-grained laws or because these implications are cognitively intractable for FISA (recall that FISA is not logically omniscient; she, like us, finds some inferences too complex to complete). So she will find another set of repeatable yet accurate generalizations at a coarser level of grain and add *these* to her lawbook,

[17] Handfield and Wilson deliver an apparatus for combining distinct objective probability functions at various levels of grain without generating the sort of contradictions described in Meacham (2014), but they do not offer a metaphysical view of probability to motivate their hierarchy. The view described here provides a motivation for the sort of hierarchical view described by Handfield and Wilson and extends the account to deterministic laws.

[18] How do we deal with indeterminate probabilities? We could just leave them undefined. But we could also go fuzzy; plausibly the best way to formally represent the FISA's epistemic state is with a set of probability functions $\mathcal{F}$ such that for each credence function $F \in \mathcal{F}$, $F(P \mid B) = 1$ if the laws predict P given B and 0 if they do not. These functions may however disagree about what credence B receives.

generating $F_2(P \mid B_2)$. When does this stop? Whenever either FISA's credal function is defined over all possible boundary conditions (unlikely) or she's unable to find laws that have an acceptable degree of both repeatability and accuracy. At that point, adding laws to the lawbook will diminish its accuracy without increasing its applicability; the FISA, wisely, will stop.

3.3 The Success of Democracy

This is the democratic view: each science represents a distinct level of grain, at which we must balance accuracy and repeatability to formulate laws. But the justification for adding new sciences is to improve the overall score of FISA's credal state in terms of accuracy, repeatability, and comprehensiveness. How, then, do we satisfy our four requirements?

METHODOLOGICAL INDEPENDENCE: I laws of each science are added to the lawbook because they individually increase the informativeness of the lawbook. Determining which generalizations will fill this role at some level of grain is the job of each special science.

COUNTERFACTUAL ROBUSTNE: the laws of each science are laws for the same reason: they increase the comprehensiveness of our system of laws without weakening its accuracy or repeatability. They support counterfactuals for the same reason the fundamental laws do. Of course, for a Humean, this story is complicated. The short version says that the laws are counterfactually robust because they ground counterfactuals by fiat; the longer version justifies this stipulation by the pragmatic utility of holding these particularly informative and supported generalizations fixed while evaluating counterfactuals.

MUTUAL CONSTRAINT: because the accuracy, comprehensiveness, and repeatability of a law system is evaluated holistically, we can expect the laws not to contradict one another—if they did so, the accuracy of the lawbook would be obviously compromised. We should expect the various sciences to inform one another. Discovering connections between sciences allows us to insure the mutual consistency of our overall belief structure.

ASYIRY: the array of facts at each level is a coarse-graining of some lower level. If B is a coarse-graining of A, then setting the value of A determines the value of B (but not vice versa). So more fined-grained information screens out more coarse-grained information, and the facts of the higher-level science are implied by the facts of the lower-level sciences. It is just this asymmetry that requires us to add special scientific laws to our system: though the fine-grained information settles the coarse-grained states, coarse-grained boundary conditions tell us nearly nothing about their fine-grained realizers. Recall that this is why we need to add higher-level laws to make predictions given coarse-grained information.

I conclude that the democratic view neatly explains all four features of the scientific coordination problem: methodological independence, lawhood, mutual constraint, and asymmetry. Its explanations of METHODOLOGICAL INDEPENDENCE and COUNTERFACTUAL ROBUSTNESS are reminiscent of the radical anarchist; its explanations of MUTUAL CONSTRAINT and ASYMMETRY are close to those of the imperialist and the moderate anarchist, respectively. In this way it poaches the best features of each of the views I've discussed.

3.4 More Fruits of Democracy

I'll now discuss some additional features of the view. First, though the view lacks the imperialism of Loewer (2009), the Past Hypothesis and PROB, from Albert (2000), still has a special status. Second, this view gives us a new way of thinking about the inexact nature of special scientific laws. Thirdly, the view correctly dispenses with Lewis's (1983) claim that special scientific terms are eligible just in case they have a simple definition in more fundamental terms. Finally, it gives us a new barometer of the lawfulness of special scientific generalizations.

Albert (2000) and Loewer (2009) argue on the basis of this sort of consideration that the lawbook must contain the Past Hypothesis (PH) and PROB, a law specifying a probability distribution over initial states. Such a distribution will yield conditional probabilities conditional on any macroscopic proposition compatible with microphysics. I've argued previously that Loewer and Albert do not go far enough because they cannot account for the lawfulness of special science generalizations; the view I defend here justifies the inclusion of the laws of the special sciences by appeal to the cognitive intractability of deriving conditional probabilities from PROB for most special science generalizations, and the fact that we cannot identify the correct constraints on these initial probabilities to get all and only the generalizations of the special sciences. However, the laws of physics together with PH and PROB together form an extremely comprehensive, repeatable, and accurate set of laws; because many other special scientific laws only apply to a very small part of the universe (right here) and are both much more inexact and have many more exceptions, adding them to the lawbook increases its comprehensibility at a greater cost to accuracy than PROB and PH. Finally, PROB and PH feature in any explanation of the existence of the complexity required to get the other special scientific generalizations going. So while I hold that the special sciences don't inherit their lawhood from physics together with PROB and PH, these laws have a special explanatory status.

This view of laws neatly accounts for another feature of special scientific laws which has been recognized by various authors (Mitchell (2000), Woodward (2003, 2013)): the laws of special sciences have exceptions, but these exceptions cannot be captured in *ceteris parabis* clauses using the concepts of the special sciences (Cartwright (1983, 1997), Woodward (2003)). On the view sketched, special scientific generalizations are lawful if and only if they feature in a system which acceptably balances accuracy, comprehensiveness, and repeatability. Laws which have exceptions can lack perfect accuracy but, by being repeatable and extending the comprehensiveness of the system, be worthwhile additions to the lawbook. Their inclusion does not require their exceptionlessness, nor does it require that there be formulated or nonredundant *ceteris parabis* conditions limiting their scope.

Thirdly, the worthiness of special science vocabulary is not dependent on its definability in fundamental terms. On Lewis's view, whether a term is eligible for use in a special science depends on its degree of naturalness; degree of naturalness depends, for Lewis, solely on the length of its definition in perfectly natural terms. This means, among other things, that the predicate 'electrino', which we stipulate to refer to electrons created before 2015 and neutrinos created after 2015, is more eligible to feature in a special scientific law than is the term 'mammal', which presumably has an extremely complex and disjunctive definition in perfectly natural terms. 'Electrino' is not more natural than 'mammal', and independently of our view of special science vocabulary we should recognize that mammals are more similar to one another than electrons are to neutrinos, whenever they are created. On the view here offered, the eligibility of a term instead depends on whether the comprehensiveness of a set of laws can be sufficiently increased by adding laws in those terms to our complete lawbook.[19]

While it is a requirement of the view that the higher-level terms force a partition of worlds which is a coarse-graining of those offered by the lower-level terms, this minimal constraint does not make the relative eligibility of coarse-grainings dependent on anything other than the informativeness of the laws so phrased, as measured by accuracy, comprehensiveness, and repeatability. These are not syntactic matters at all. So the apparent need for a metaphysically special language to evaluate the lawfulness of the special sciences is significantly ameliorated.

Finally, laws come with various degrees of lawfulness. Some laws are less modally robust: they hold in fewer situations, and they are less stable than others. There is a continuum of laws, starting with the laws of physics, which are exceptionless and maximally modally robust, moving through the central principles of special

[19] For a more in-depth discussion of the difficulties involved in tying the Lewisian notion of naturalness to our account of laws, see Loewer (2007) and Eddon and Meacham (2013); for a discussion of this problem focusing on special scientific laws, see Callender and Cohen (2009).

sciences, like the principle of natural selection or the thermal relaxation time of a certain sort of liquid, and culminating in mere accidental generalizations. The view sketched, unlike more modally robust accounts of laws, has the capacity to account for this. For there is more than one way to weight the three virtues this view rests on: if perfect (or near-perfect) accuracy is given maximal weight, then only the laws of physics are included in the lawbook. By varying our permissiveness for accuracy, we will vary the generalizations which are permitted in the lawbook. Those which count as laws on more accurate rankings occupy a more privileged place on this continuum than those which do not.

This hierarchy can be tied to the counterfactual robustness of the laws. For the view sketched is Humean, according to which counterfactuals are made true by the laws. Plausibly,[20] the strength of counterfactual support varies with the accuracy of the laws.[21] So counterfactuals which are made true by the laws of physics, our most accurate set, override those made true by biology. And within a science, the counterfactuals made true by more accurate laws trump those made true by less accurate laws—so the counterfactuals made true by quantum mechanics trump those made true by classical mechanics.

I've claimed that a particular form of the Best Systems Account of lawhood can explain the relevant features of the relationship between laws in various scientific disciplines. Can a more metaphysically robust view do this? I am doubtful. For a key feature of the view is taking the informativeness of the lawbook, measured in a particular way, to be partially constitutive of lawhood. Anti-Humeans reject the claim that the laws are, by their nature, informative. So no way of measuring the informativeness of the lawbook will suffice to make some higher-level generalization a law.

Nonetheless proponents of metaphysically robust views who hold that *only* the fundamental laws are backed by modally robust fundamental facts can appeal to the view I've defended to distinguish between accidents and laws at a higher level. Many philosophers who doubt that a fully Humean story can be told about the fundamental structure of the world are subject to the criticisms laid at the feet of the imperialist in Section 2.1. Although the view that results will have a different explanation of the counterfactual robustness of the special science laws than that of the laws of physics, they will inherit the other advantages of the democratic view.

[20] Or by stipulation!

[21] Woodward (2003), Section 6.12, argues against Sandra Mitchell's (2000) notion of *stability* and Brian Skyrms's (1977) notion of *resiliency* on the basis that these non-modal notions cannot capture what we are really interested in in discovering laws, vis, their counterfactual stability (this point is also made by Lange (2009)). On the view offered, as in other Humean views, counterfactual stability is grounded in occurrent facts, in this case, a sort of stability across situations in a similar vein to that described by Mitchell and Skyrms. So it would be a mistake to criticize this view for missing the counterfacts—they are true because of the occurrent facts described. Of course, the proof is in the pudding: does the sort of stability here described generate the *right* counterfactuals? I hold that it does.

4 Conclusion

Extant views describing the relationship between distinct scientific disciplines leave key features of this relationship unexplained. This failure manifests itself in philosophical views about the lawhood of special scientific laws; these views, no matter their metaphysical commitments, fail either to account for the autonomy of the special sciences or for the mutual dependence of scientific disciplines. I have here shown that a Humean view, which takes the informativeness of the laws to be partially constitutive of their lawhood, measures informativeness by the accuracy of predictions made by the laws on the basis of repeatable boundary conditions, and evaluates the informativeness of all sciences together, is uniquely able to capture these features of the relationship between laws. I then pointed out additional advantages of the view: it accounts for the degrees of lawfulness of special scientific laws, the fact that special scientific laws have exceptions, and the fact that explaining these exceptions often requires concepts that are not a part of the science in which the law is formulated.

References

Adlam, Emily (2022). Laws of Nature as Constraints. *Foundations of Physics* 52 (1): 1–41

Albert, David (2000). *Time and Chance*. Cambridge, MA: Harvard University Press.

Armstrong, David (1983). *What Is a Law of Nature?* Cambridge: Cambridge University Press.

Armstrong, David (1997). *A World of States of Affairs*. Cambridge: Cambridge University Press.

Batterman, Robert W. (2001). *The Devil in the Details: Asymptotic Reasoning in Explanation, Reduction and Emergence*. Oxford: Oxford University Press.

Beatty, John (1995). "The Evolutionary Contingency Thesis." In Gereon Wolters and James G. Lennox (eds), *Concepts, Theories, and Rationality in the Biological Sciences*, 45–81. Pittsburgh: University of Pittsburgh Press.

Beebee, Helen (2000). "The Non-Governing Conception of Laws of Nature." *Philosophy and Phenomenological Research* 61: 571–594.

Beebee, Helen (2006). "Does Anything Hold the Universe Together?" *Synthese* 149 (3): 509–533.

Beebee, Helen (2011). "Necessary Connections and the Problem of Induction." *Nous* 45 (3): 504–527.

Bird, Alexander (2007). *Nature's Metaphysics: Laws and Properties*. Oxford: Oxford University Press.

Butterfield, Jeremy (2011). "Less is Different: Emergence and Reduction Reconciled." *Foundations of Physics* 41: 1065–1135.

Callender, Craig, and Jonathan Cohen (2009). "A Better Best System Account of Lawhood." *Philosophical Studies* 145 (1): 1–34.

Callender, Craig, and Jonathan Cohen (2010). "Special Science, Conspiracy, and the Better Best System Account of Lawhood." *Erkenntnis* 73: 427–447

Cartwright, Nancy (1983). *How The Laws of Physics Lie*. Oxford: Oxford University Press.

Cartwright, Nancy (1997). "Where do Laws of Nature Come From?" *Dialectica* 51 (1): 65–78.

Cartwright, Nancy (1999). *The Dappled World: A Study at the Boundaries of Science*. Cambridge: Cambridge University Press.

Chen, Eddy Keming, and Sheldon Goldstein (2022). "Governing Without a Fundamental Direction of Time: Minimal Primitivism about Laws of Nature". In Yemima Ben-Menahem (ed.), *Rethinking the Concept of Law of Nature*, 21–64. Cham: Springer.

Eddon, Maya, and Chris Meacham (2013). "No Work for a Theory of Universals." In Barry Loewer and Jonathan Schaffer (eds.), *A Companion to David Lewis*, 116–137. Blackwell.

Ellis, Brian (2001). *Scientific Essentialism*. Cambridge University Press.

Fenton-Glynn, Luke. (2016). "*Ceteris Paribus* Laws and *Minutis Rectis* Laws." *Philosophy and Phenomenological Research* 93 (2): 274–305.

Fodor, Jerry (1974). "The Disunity of Science as a Working Hypothesis." *Synthese* 28: 97–115.

Frisch, Mathias (2014). "Why Physics Can't Explain Everything." In Alastair Wilson (ed.), *Chance and Temporal Asymmetry*, 221–240. Oxford University Press.

Handfield, Toby, and Wilson, Alastair (2013). "Chance and Context." In Alastair Wilson (ed.) *Chance and Temporal Asymmetry*, 19–44. Oxford University Press.

Hall, Edward (2015). "Humean Reductionism about Laws of Nature." In Barry Loewer and Jonathan Schaffer (eds), *A Companion to David Lewis*, 262–277. Oxford: John Wiley & Sons, Ltd.

Hicks, Michael Townsen (2017) "Dynamic Humeanism." *British Journal for the Philosophy of Science* 69: 983–1007.

Hoefer, Carl (2007). "The Third Way on Objective Probability: A Sceptic's Guide to Objective Chance." *Mind* 116 (463): 549–596.

Hoefer, Carl (2014). "Consistency and Admissibility: Reply to Meacham." In Alastair Wilson (ed.), *Chance and Temporal Asymmetry*, 68–80. Oxford University Press.

Hüttemann, Andreas (2014). "Ceteris Paribus Laws in Physics." *Erkenntnis* 79 (S10): 1715–1728.

Lange, Marc (2009). *Laws and Lawmakers: Science, Metaphysics, and the Laws of Nature*. Oxford: Oxford University Press.

Lewis, David (1980). "A Subjectivist's Guide to Objective Chance." In Richard C. Jeffrey (ed.), *Studies in Probability and Inductive Logic*, 263–293. University of California Press.

Lewis, David (1983). "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61 (4): 343–377.

Loewer, Barry (2007). "Laws and Natural Properties." *Philosophical Topics* 35 (1/2): 313–328.

Loewer, Barry (2008). "Why There Is Anything Except Physics." In Jakob Hohwy and Jesper Kallestrup (eds), *Being Reduced: New Essays on Reduction, Explanation, and Causation*, 149–163. Oxford: Oxford University Press.

Loewer, Barry (2009). "Why Is There Anything Except Physics?" *Synthese* 170 (2): 217–233.

Maudlin, Tim (2007). *The Metaphysics Within Physics*. Oxford: Oxford University Press.

McKenzie, Kerry (Forthcoming). "In No Categorical Terms: A Sketch for an Alternative Route to Humeanism about Fundamental Laws." In M. C. Galavotti, S. Hartmann, M. Weber, W. Gonzalez, D. Dieks, and T. Uebel (eds), *New Directions in the Philosophy of Science*, 45–61. New York: Springer.

Meacham, Christopher J. G. (2013). "Autonomous Chances and the Conflicts Problem." In Alastair Wilson (ed.), *Asymmetries in Chance and Time*, 45–67. Oxford University Press.

Mitchell, Sandy (2000). "Dimensions of Scientific Law." *Philosophy of Science* 67 (2): 242–265.

Ramsey, Frank Plumpton (1929). "General Propositions and Causality." Reprint. In D. H. Mellor (ed.) *F. P. Ramsey: Philosophical Papers*, 34–51. Cambridge: Cambridge University Press, 1990.

Schrenk, Markus (2008). "A Lewisian Theory for Special Science Laws." In H. Bohse and S. Walter (eds.), *Selected Contributions to GAP.6, Sixth International Conference of the Society for Analytic Philosophy. Berlin, 11–14 September 2006*, 121–131. Paderborn: Mentis.

Schrenk, Markus (2014). "Better Best Systems and the Issue of CP-Laws." *Erkenntnis* 79 (10): 1787–1799.

Skyrms, Brian (1977). "Resiliency, Propensities, and Causal Necessity." *Journal of Philosophy* 74 (11): 704–713.

Weslake, Brad (2014). "Statistical Mechanical Imperialism." In Alastair Wilson (ed.), *Chance and Temporal Asymmetry*, 241–257. Oxford University Press.

Wilson, Alastair (Forthcoming). "Metaphysical Emergence as Higher-Level Naturalness." In David Yates (ed.), *Rethinking Emergence*, Oxford University Press.

Woodward, James (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, James (2013). "Laws, Causes, and Invariance." In Stephen Mumford and Matthew Tugby (eds), *Metaphysics and Science*, 48–72. Oxford: Oxford University Press.

Woodward, James (2014). "Simplicity in the Best Systems Account of Laws of Nature." *British Journal for the Philosophy of Science* 65 (1): 91–123.

Index

For the benefit of digital users, indexed terms that span two pages (e.g., 52–53) may, on occasion, appear on only one of those pages.

Tables and figures are indicated by an italic *t* and *f* following the page number.