Jerusalem Studies in Philosophy and History of Science

Stavros Ioannidis
Gal Vishne
Meir Hemmo
Orly Shenker *Editors*

Levels of Reality in Science and Philosophy

Re-examining the Multi-level Structure of Reality

Sidney M. Edelstein Center
for the History and Philosophy
of Science, Technology and Medicine

Springer

Jerusalem Studies in Philosophy and History of Science

Jerusalem Studies in Philosophy and History of Science sets out to present state of the art research in a variety of thematic issues related to the fields of Philosophy of Science, History of Science, and Philosophy of Language and Linguistics in their relation to science, stemming from research activities in Israel and the near region and especially the fruits of collaborations between Israeli, regional and visiting scholars.

Stavros Ioannidis • Gal Vishne • Meir Hemmo •
Orly Shenker
Editors

Levels of Reality in Science and Philosophy

Re-examining the Multi-level Structure of Reality

Springer

Editors
Stavros Ioannidis
Department of History & Philosophy of Science
National & Kapodistrian University of Athens
Athens, Greece

Meir Hemmo
Philosophy Department
University of Haifa
Haifa, Israel

Gal Vishne
Edmond and Lily Safra Center for Brain Sciences
The Hebrew University of Jerusalem
Jerusalem, Israel

Orly Shenker
Sidney M. Edelstein Center for the History & Philosophy of Science
Technology and Medicine The Hebrew University of Jerusalem
Jerusalem, Israel

ISSN 2524-4248 ISSN 2524-4256 (electronic)
Jerusalem Studies in Philosophy and History of Science
ISBN 978-3-030-99424-2 ISBN 978-3-030-99425-9 (eBook)
https://doi.org/10.1007/978-3-030-99425-9

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Contributors

Yemima Ben-Menahem The Hebrew University of Jerusalem, Jerusalem, Israel

Sean M. Carroll California Institute of Technology, Pasadena, CA, USA

Michael Esfeld University of Lausanne, Lausanne, Switzerland

Erez Firt University of Haifa, Haifa, Israel

Nir Fresco Ben-Gurion University of the Negev, Be'er Sheva, Israel

John Heil Washington University in St. Louis, St. Louis, MO, USA
Durham University, Durham, UK
Monash University, Melbourne, VIC, Australia

Meir Hemmo University of Haifa, Haifa, Israel

Stavros Ioannidis National & Kapodistrian University of Athens, Athens, Greece

Christian List Ludwig-Maximilian University of Munich, Munich, Germany
London School of Economics, London, UK

Gualtiero Piccinini University of Missouri – St. Louis, St. Louis, MO, USA

Stathis Psillos National & Kapodistrian University of Athens, Athens, Greece

Lawrence Shapiro University of Wisconsin – Madison, Madison, WI, USA

Orly Shenker The Hebrew University of Jerusalem, Jerusalem, Israel

Elliott Sober University of Wisconsin – Madison, Madison, WI, USA

Gal Vishne The Hebrew University of Jerusalem, Jerusalem, Israel

James Woodward University of Pittsburgh, Pittsburgh, PA, USA

Chapter 1
Introduction

Stavros Ioannidis, Gal Vishne, Meir Hemmo, and Orly Shenker

Abstract In this introductory chapter we present some central philosophical views and problems about the notion of levels of reality that will be further explored in the chapters of this volume. We point out that the question whether reality has a multi-level structure is a deep philosophical issue with widespread implications for how we think about central problems in philosophy and science. We emphasise the many aspects of the notion of levels, distinguish between ontological levels (where levels are used as a way to talk about the hierarchical structure of the world) and epistemological levels (where levels have primarily a methodological and epistemological role) and explore their complex relationship. We also discuss the general reasons offered by non-reductive physicalists to adopt a metaphysics of a multi-level reality and whether the levels described by such accounts of the special sciences can be part of a physicalist ontology.

1.1 Levels of Reality in Science and Philosophy

Does reality contain many levels or is the world 'flat', in the sense that everything is fully reducible to some fundamental level? We take this to be one of the deepest questions about the world we live in with widespread implications for both science and philosophy.

The view that reality has a multi-level structure is an idea that may seem hard to deny, given contemporary science; and yet (as we shall see below and in the chapters

S. Ioannidis (✉)
National & Kapodistrian University of Athens, Athens, Greece
e-mail: sioannidis@phs.uoa.gr

G. Vishne · O. Shenker
The Hebrew University of Jerusalem, Jerusalem, Israel

M. Hemmo
University of Haifa, Haifa, Israel

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_1

of this volume), it has been the source of deep philosophical puzzles. Science has revealed a world that contains many kinds of things, from electrons to organisms and societies. The language of levels has been commonly used in science and philosophy to refer to this diversity, the underlying idea being that things form a kind of a hierarchy, with entities at lower-levels of the hierarchy composing entities at higher levels. The scientific image, thus, seems to reveal a reality with a hierarchy of levels.

The notion of levels has epistemological and methodological aspects too. So, different scientific fields have been thought to correspond to different levels of this hierarchy (Oppenheim & Putnam, 1958). Moreover, it is commonly said that a single phenomenon can be approached and explained via different levels of analysis or explanation. In biology, for example, phenomena can be investigated at molecular, cellular, or organismic levels (e.g. limb development or the pathology of a disease). In brain and cognitive sciences, in particular, talk about levels is literally built into the subject matter of investigation, which is the relationship between, on the one hand, the workings of the brain (at various 'levels') and, on the other hand, cognition and behaviour. As these examples show, the notion of levels is central in scientific practice.

Epistemological and ontological aspects of levels talk are intertwined. In particular, if one takes a realistic view about our best scientific theories, the question: 'Are there levels of reality?' cannot be avoided. We take it that for a realist there should be something to which one refers to when one takes seriously the idea that, for example, biology and physics are theories of the world at different levels, or that phenomena (in life sciences and elsewhere) can typically be investigated at various levels. Since levels seem to be central in a scientifically oriented point of view, one has to explain what exactly they are and how they are connected. Unless one wishes to take a more instrumental view with respect to the notion of levels (as mere tools of explanation, or description—see below), a realistic attitude towards our scientific theories seems to commit us to a multi-level structure of reality.

The centrality of the notion of levels in scientific practice combined with a realist attitude towards theories, together with the familiarity of a hierarchically structured scientific image, makes a multi-level ontology very appealing. But apart from its centrality in science, the notion of levels has been important in philosophical debates too. Two main such examples of philosophical discussions where intuitions about levels have been decisive are the issue of the relation between the so-called special sciences and physics, and the mind-body problem.

Almost everybody agrees that we are made of particles or matter fields, but almost nobody thinks that this is the end of the story, for example, that our mental states are nothing but configurations of particles or fields (or whatever current and future physics tells us the world is made of). In the case of the relation between brains and minds, then, we commonly think of minds and mental properties as being 'at a higher level' than brains. Importantly, using the notion of levels to form intuitions about the mind-body problem is not only confined within philosophy. Psychologists, cognitive scientists and neuroscientists investigating such issues as the precise relation of the mental to the underlying neural structures and the appropriate approach to study it, are confronted with such questions as whether the mental

is physical or whether it exists at a 'higher' computational 'level' (this example illustrates that far from being an abstract philosophical problem, contemporary major research programs in brain and cognitive sciences are determined by implicit views concerning the mind-body problem).

When philosophers think about the status of special sciences and whether they are autonomous from physics, intuitions about levels are again central. Entities postulated by special sciences and special science natural kinds, such as cells, organisms, psychological states and societies, are thought to be at a 'higher level' than the level of molecules, particles and fields. Physics, then, is taken to describe the 'fundamental level' of reality (or possibly just one deep level in a non-foundationalist picture), and sciences such as chemistry, biology and psychology are taken to be about 'higher levels'. This picture is adopted by non-reductive physicalism, an extremely influential view, according to which it may be the case that everything is in some sense physical, and yet there is something more at higher levels (we will come back to non-reductive physicalism below).

1.2 Epistemological vs. Ontological Levels

We have said that a multi-level reality gives rise to deep philosophical puzzles. In the remaining of this introduction we will discuss some of them, focusing on two main questions: What exactly are we ontologically committed to in accepting a multi-level structure of reality? And how satisfactory are the arguments in its favour?

We have already noted that the prevalence of levels talk in science may be thought to justify a multi-level ontology (in contrast to a 'flat' one). How justified is this inference? A problem here is that the notion of levels is used for various purposes and in various contexts in science and philosophy (see also Craver, 2015); its exact content is thus context-dependent. It is therefore important to distinguish between the various uses of the notion. In particular, it is important to distinguish between two ways to think about levels. We can think about levels as primarily an item in methodology and explanation; and we can use the concept of levels as primarily a metaphysical notion, i.e. as a way to talk about the hierarchical structure of the world.

Both ways to think about levels are to be found in influential contemporary discussions. For example, Craver's (2007) view of 'levels of mechanisms' takes levels as primarily an epistemological concept, important when constructing mechanistic explanations of phenomena (for levels of explanation see also Woodward, Chap. 14, this volume). Wimsatt's well-known account of levels of organisation as "local maxima of regularity and predictability" (1976, 209) is similarly science-based, but Wimsatt is explicit that his notion of levels corresponds to ontological features of the world. Both these accounts of levels can be contrasted with discussions (e.g. in the context of non-reductive physicalism—see below) that take levels as primarily a notion important in metaphysics.

What are the relationships between the two ways to think about levels? There are two possibilities: either the two senses are related in the sense that levels in science imply ontological levels; or they have to be sharply distinguished, in the sense that there is no implication from epistemological to ontological conclusions. That such a relation exists, is a common view among, for example, philosophers of biology and of neuroscience. For philosophers who adopt such a perspective, the notion of levels is important both in methodology and epistemology and in ontology. For such philosophers, levels of explanation or description (epistemological levels) correspond to real levels in nature: our world is hierarchically structured. The underlying thought here is that if it is true that the notion of levels is central in science, then this points to a multi-level ontological view of reality. Conversely, if nature contains many levels, then this will have to be reflected in scientific practice; since this seems to be exactly what we find in science, some version of a multi-level ontology has to be correct.

But it is also possible to argue that one has to separate epistemological and ontological senses of levels. That is, one could remain non-committal about metaphysics, and explore instead how the notion of levels functions in scientific practice. According to such a perspective, levels are important as a methodological and epistemological item of scientific practice, but not as an ontological feature. Levels talk in biology and neuroscience, for example, can be viewed as a way to organise research practice and coordinate explanations from various domains, but does not necessarily lead to postulating an ontological hierarchy of levels. On such a view, a scientific practice where the notion of levels plays an important role is in principle compatible with both a multi-level reality and a flat one.

The difference between these two perspectives can be illustrated by the example of 'mechanistic levels' in mechanistic explanations (see Craver, 2007, Ioannidis & Psillos, Chap. 9, this volume). Mechanists that are more interested in how mechanisms function within scientific practice are less inclined to take mechanistic levels in a robust metaphysical sense. One may endorse a hierarchy of mechanistic levels as a way to systematise how mechanistic explanations in biology are constructed, how different lines of research are coordinated, etc., without being committed to a comprehensive multi-level ontology. Other mechanists (e.g. Glennan, 2017) are interested in developing a systematic metaphysics based on mechanisms; such mechanists may be more inclined to interpret mechanistic hierarchies in robust metaphysical terms.

Philosophers mainly interested in metaphysics may also adopt the view that ontological levels and epistemological levels are to be kept separate. Such philosophers may want to argue for a multi-level ontology, or alternatively for an ontology without levels, for reasons other than how the notion of levels functions within science (for accounts of ontologies without levels, see Esfeld, Chap. 4, this volume, Heil (2003) and Chap. 7, this volume, Hemmo & Shenker, Chap. 8, this volume). For example, a common motivation is clarifying the relation between the mental and the physical (however, many philosophers who adopt a multi-level ontology think of epistemological and ontological levels as closely related). Ontological views that reject a hierarchy of levels need to account for why the notion of levels seems central

to scientific practice and explain whether in such a view the autonomy of special sciences can be preserved.

The relationship between epistemological and ontological aspects of levels talk has implications for whether we are inclined to accept a multi-level ontology and so for several central problems in philosophy of mind, philosophy of science and metaphysics. What further complicates the picture (as well as the possible inference from levels talk in science to ontological conclusions) is the diversity of level concepts that we find in science and philosophy. Let us briefly examine some of them.

1.2.1 Ontological Levels: The Layered Model and Its Alternatives

One can think about levels of reality in a 'minimal' sense, e.g. when entities that compose or are parts of other entities are described as being at a 'lower' level. But often one thinks about levels in a more robust sense. For example, Kim has written about the 'layered model' of the world, by which he meant "a single hierarchy of connected levels, from higher to lower, in which every object and phenomenon of the natural world finds its "appropriate" place" (2002, 16). Such a structure is for example taken by Oppenheim and Putnam (1958) to underlie the unity of science. In the Oppenheim and Putnam version, this 'layer cake' model is taken to imply that for any two objects, either one is higher than the other, or they are both on the same level; moreover, that entities at level n can only be composed of entities at the directly lower-level $n-1$. A different version of the layered model was put forward by the British Emergentists (e.g. Morgan, 1923), where, as Kim (2002) notes, what generates the hierarchical structure are not part-whole relations, as in Oppenheim and Putnam, but relations of emergence.

Such stronger views of 'levels of reality' have been shown to lead to problems (see Kim, 2002, Craver, 2007, Potochnik & McGill, 2012 and Shapiro, Chap. 12, this volume). Potochnik and McGill, for example, have argued that the notion of levels (of composition) presupposes that "atoms must always compose molecules, populations must always compose communities, and so forth"; however and by contrast, such "uniformity of composition needed for stratified levels simply does not exist" (2012, 126). It is possible nevertheless to obtain hierarchical structures by modifying these stronger assumptions; for example, by viewing levels as forming a tree-like structure rather than a linear hierarchy (Wimsatt, 1976), or by viewing levels as local rather than global (as in levels of mechanisms, see Craver, 2007). The rejection of the layered model, thus, does not lead to the rejection of all kinds of levels of reality. Moreover, some kind of levels hierarchy is (arguably) presupposed by the debate of (ontological) reductionism vs. anti-reductionism, since

reductionism (or anti-reductionism) is commonly construed as the view that higher levels can be reduced (or cannot be reduced) to lower-levels.[1]

1.2.2 Epistemological Levels: Pluralism and Skepticism

When we look more closely at scientific practice, what we find is not a specific notion of levels, but a family of various notions, more or less loosely connected to the metaphor of 'levels'. Scientists talk, for example, about levels of abstraction, analysis, causation, description, complexity, explanation, processing and organisation, among others. This plurality of conceptions gives rise to a natural question: is there a single concept of levels, or is there instead a plurality of distinct notions? Answering this question is important for both perspectives identified above. On the one hand, to clarify the work that the notion of levels does as a methodological and epistemological item, we have to take into account this diversity of uses. On the other hand, clarifying how the notion is used in science and how different uses connect to each other, is crucial for thinking more clearly about the ontological commitments of levels talk.

The diversity of level concepts in science and the difficulties with some central ideas associated with levels talk have prompted some philosophers to question the usefulness of the notion (see Potochnik & McGill, 2012). Skeptics about levels have adopted a 'deflationary' approach (see Eronen, 2015), suggesting that the notion of levels should give way to other, better-defined notions, such as scale or composition. Such 'levels skepticism' (cf. Eronen & Brooks, 2018) casts doubt on the extent to which the notion of levels is required to make sense of scientific practice. Skepticism about levels has implications for the view of a multi-level reality too: if levels are not really a feature of the scientific image or a central item of scientific practice, then the idea that there are (ontologically speaking) different levels of reality needs to be reconsidered.

We see thus that there are different kinds of levels in science and philosophy. We have distinguished between epistemological and ontological levels and discussed some challenges for the inference from levels talk in science to a multi-level ontology. Let us now explore some more general reasons to adopt a multi-level ontology, that have been offered in the context of non-reductive physicalism.

[1] How exactly levels are construed is of course crucial for the reductionism vs. anti-reductionism debate. Thus, Oppenheim and Putnam's layered model has been linked to a reductionist account, whereas levels of mechanisms, as well as levels of complexity and organisation, are connected to a broadly anti-reductionist attitude.

1.3 Non-reductive Physicalism as a Multi-level View

During the second half of the twentieth century physicalist thinking has become central in analytic philosophy. Contemporary physicalism contains two central ideas. On the one hand, physicalists stress the primacy of physics for describing the fundamental ontology of the world: everything is ultimately physical. On the other hand, this does not mean that physical facts are *all* there is. According to the dominant version of physicalism that has been called non-reductive physicalism, some 'higher-level' facts, described by the so-called special sciences (e.g. biology or psychology), cannot in principle be reduced to the fundamental physical level. The main intuition supporting this idea is that special science kinds are multiply realisable by physical kinds: namely, the intuition is that the same higher-level kind may be realised by physical tokens that do not share any (relevant) physical property, and in this sense they do not belong to the same (relevant) physical kind. So, although the physical facts determine all the facts, according to non-reductive views, there is something about higher levels that is not completely fixed by the physical facts. In the language of supervenience (where high-level facts are taken to supervene on low-level ones), given a high-level fact, the entirety of all the physical facts do not fully determine the set of physical kinds that forms its supervenience basis. This idea of irreducibility has been expressed very clearly in different ways by many philosophers (e.g., Putnam, 1967/1975; Davidson, 1970; Fodor, 1974). In one way or another, all non-reductive physicalists accept the idea that reality consists of different levels in the sense that higher-level facts are not reducible to the facts at other (typically) lower levels.

To better understand what it is to have a multi-level view of reality, let us consider the notion of supervenience in some more detail. Non-reductive physicalists accept the idea that high-level kinds supervene on lower-level kinds and ultimately on physical kinds. By supervenience here, one means that there can be no change in a high-level kind without some change in the physical kind of the realiser. (Compare this with the idea of multiple realisability according to which there can be a change in physical kind that does not require a change in the mental kind it realises and the amount of freedom or independence it leaves for the higher level). This idea of supervenience is taken to be the hallmark of physicalism since it seems to guarantee some sort of dependence of the higher levels on the physical level: in some sense it implies that the facts (or kinds) at the physical level fix or determine the facts (or kinds) at higher levels. But the dependence here is quite weak since the details are left open. Supervenience is a formal relation (between kinds, facts, properties etc.) and as such it is compatible with a variety of metaphysical relations—even with reductive type-type physicalism: if there is a 1:1 relation (instead of 1:many) between higher and lower-level kinds, this is compatible with taking higher-level kinds to be type-identical with lower-level kinds. So, the non-reductive physicalist needs to explain what is the specific metaphysical account that underlies the supervenience relation between kinds (or other entities), giving rise to a multi-level structure. Different views of the nature of the metaphysical facts that

underlie the supervenience relation give rise to different versions of non-reductive physicalism.

The idea that reality contains many levels, even if the supervenience relation is satisfied, is thus compatible with various kinds of metaphysical theories. The common feature of all such views is a rejection of the claim that higher-level facts described in sciences such as biology and psychology are identical with physical facts (as espoused by type-identity reductive physicalism). But there are many ways to cash out the exact nature of the relation between levels. For example, some non-reductive physicalists focus on realisation (e.g. Aizawa & Gillett, 2009; Polger & Shapiro, 2016, Shapiro, Chap. 12, this volume), while others take grounding to be the important metaphysical relation (see Tahko & Lowe, 2020). Note also that relations between higher and lower-level facts need not be 1:many; they can also be 1:1. In the latter case, the higher level can still be thought to be realised by the lower one (cf. Polger & Shapiro, 2016), without taking the two to be identical (alternatively, a 1:1 relation is compatible with certain theories of grounding).[2]

However, the most popular metaphysical account of a multi-level ontology has been the version of non-reductive physicalism that endorses multiple realisability (which is supposed to explain the irreducibility of the higher to the lower level). Reductive physicalism, in particular, according to which there is only one level and all phenomena and regularities described by the special sciences can be explained in terms of it, has become a minority view. The non-reductive version of physicalism enables one to hold on to the traditional materialist thesis that everything is at bottom physical, while at the same time viewing higher-level facts as irreducible (as a matter of principle or law) to the lower level. In that way, it is taken to guarantee also the autonomy of the special sciences (Fodor (1974), for example, has emphasised this point). In the case of psychology, in particular, a multi-level metaphysics seems to secure rationality and freedom (as emphasised by Davidson's (1970) anomalous monism), but also the special nature of the mental (e.g. the nature of qualia and the so-called hard problem of consciousness; see also Bennett, 2011).

Non-reductive metaphysical theories that accept multiple realisability have to explain what unifies a set of heterogeneous lower-level kinds into one high-level kind. In answering this difficulty, Putnam's (1967/1975) idea of a common computational-functional role shared by such lower-level kinds, Davidson's (1970) idea of sameness under a description and the idea of a common causal-functional role proved very influential. But as Fodor (1974, 1997) has observed, from a physical point of view the grouping of lower-level kinds in a higher-level one still seems like a brute fact: while higher functional kinds seem "nomologically homogeneous under their functional description" (1997, 153), there is no explanation why only

[2] A multi-level ontology need not be only part of a physicalist account, as the exact nature of the lower-level facts does not matter for whether one is committed to a hierarchy of levels. Dualists that take the fundamental level to be both physical and mental, for example, can also adopt a multi-level ontology.

certain lower-level kinds, and not others, fall under a given higher-level kind. As Fodor expresses this point, which he describes as "molto mysterioso":

> Only God ... gets to decide whether there are laws about pains; or whether, if there are, the pains that the laws are about are MR [multiply realized]. (Fodor, 1997, 161)

A central feature of multi-level ontologies is that the non-reducible higher-level facts or kinds that they posit are not inert or superfluous, but feature in laws of nature and/or are thought to be causally efficacious. The supposed causal efficacy of higher-level properties, in particular, gives rise to a central objection to multi-level ontologies, i.e. that since physical effects have sufficient physical causes, higher-level causation leads to overdetermination and is therefore to be rejected. But this line of argument too is inconclusive.

The chapters in the present volume reconsider the view that reality contains many levels and open new ways to understand the status of the special sciences, with special emphasis on physics and the physical-mental relation. They present state-of-art research on these problems and discuss various aspects of the conception of levels of reality, emphasising the contribution of science to the philosophical discussion and vice versa. Although epistemological aspects of the notion of levels will be examined in several of the chapters, the main focus will be on the metaphysics of a multi-level reality, on whether the levels described by various non-reductive accounts of the special sciences can be part of a physicalist ontology, and on exploring 'flat reality' alternatives. We would like to thank all reviewers of the chapters for their kind help and valuable feedback they provided to the authors of this volume.

References

Aizawa, K., & Gillett, C. (2009). The (multiple) realization of psychological and other properties in the sciences. *Mind and Language, 24*, 181–208.

Bennett, K. (2011). By our bootstraps. *Philosophical Perspectives, 25*, 27–41.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Craver, C. F. (2015). Levels. In T. Metzinger & J. M. Windt (Eds.), *Open MIND* (pp. 1–26). MIND Group. https://doi.org/10.15502/9783958570498

Davidson, D. (1970). Mental events. In L. Foster & J. W. Swanson (Eds.), *Experience and theory* (pp. 207–224). Duckworth.

Eronen, M. I. (2015). Levels of organization: A deflationary account. *Biology and Philosophy, 30*, 39–58.

Eronen, M. I., & Brooks, D. S. (2018). Levels of organization in biology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018 ed.) https://plato.stanford.edu/archives/spr2018/entries/levels-org-biology/

Fodor, J. (1974). Special sciences (Or: The disunity of science as a working hypothesis). *Synthese, 28*, 97–115.

Fodor, J. (1997). Special sciences: Still autonomous after all these years. *Nous, 31*, 149–163.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Heil, J. (2003). Levels of reality. *Ratio, 16*, 205–221.

Kim, J. (2002). The layered model: Metaphysical considerations. *Philosophical Explorations, 5*, 2–20.

Morgan, C. L. (1923). *Emergent Evolution*. Williams and Norgate.

Oppenheim, P., & Putnam, H. (1958). The unity of science as a working hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories, and the mind-body problem, Minnesota studies in the philosophy of science II* (pp. 3–36). University of Minnesota Press.

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford University Press.

Potochnik, A., & McGill, B. (2012). The limitations of hierarchical organization. *Philosophy of Science, 79*, 120–140.

Putnam, H. (Ed.). (1967/1975). The nature of mental states. In *Mind, language and reality: Philosophical papers* (Vol. 2, pp. 429–440). Cambridge University Press.

Tahko, T. E., & Lowe, E. J. (2020). Ontological dependence. In E. N. Zalta (Ed.), *The Stanford encyclopedia of Philosophy* (Fall 2020 ed.) https://plato.stanford.edu/archives/fall2020/entries/dependence-ontological/

Wimsatt, W. C. (1976). Reductionism, levels of organization, and the mind–body problem. In G. Globus, I. Savodnik, & G. Maxwell (Eds.), *Consciousness and the brain* (pp. 199–267). Plenum Press.

Chapter 2
Levels of Reality and Levels of Description

Yemima Ben-Menahem

In Memory of Margie Morrison

Abstract The assumption of the causal closure of the fundamental level of reality has been used to support reductionism and undermine non-reductive views such as Davidson's anomalous monism. Jaegwon Kim, in particular, devoted numerous papers to this line of critique, arguing that the stratification of reality into distinct levels is incompatible with the causal closure assumption. Taking issue with Kim's position, my chapter seeks to show that the stratified picture is both safe and useful from the scientific point of view. The defense of non-reductive physicalism requires a clear distinction between levels of reality and levels of description, a distinction that counter-arguments (such as Kim's) tend to blur.

2.1 Introduction

A few years ago I was re-described by Orly Shenker and Meir Hemmo as a dualist. At first I was taken aback; my view and Descartes' didn't seem to me to have much in common, and besides, it didn't exactly seem like a compliment . . . But then, I thought, there is no point in arguing about names. So let me accept my description as a dualist and see what it involves.

My description as a dualist, according to Shenker and Hemmo, picks out an aspect of me. I am a woman, a parent, a philosopher, an Israeli, a person whose family name begins with 'B.' These aspects are picked out by various descriptions of me and so is dualism. These aspects, moreover, are supposed to be aspects of my *physical* state, or partial descriptions of this state. Thus far aspect language is not controversial. But here we arrive at a juncture that leads in two different directions.

Y. Ben-Menahem (✉)
The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: yemima.ben-menahem@mail.huji.ac.il

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_2

It could be simply the case, simply what aspect-jargon means, that as it happens, my physical state at a particular moment satisfies the above descriptions—it is a physical state that could be described as a person whose name begins with a 'B,' or a physical state (that manifests a neurological state) that could be described as belief in dualism. Let us call this option the D option, to remind us of Davidson. The various descriptions, in this case, may be applied to an individual physical state—a token—and point to the membership of this token in a multitude of different sets, the set of dualists, the set of persons whose name begins with a B and so on. In the D option—this is the crux of the matter—there is no guarantee that these various sets constitute physical types, and in many cases, for example the set of B-named persons, it would be reasonable to deny that they do. The denial implies multiple realization, namely, it implies that descriptions such as 'dualist', or 'B-named person' could be realized by physical states that do not generate a physical type. The D option is the basis of Davidson's account of the mental, anomalous monism as he calls it (1980b), or nonreductive physicalism as it is also commonly referred to.

Aspect language, however, could also indicate—and this is presumably how full blown reductive physicalism has it—that this physical state of me instantiates a physical *type,* a type of physical state shared by all persons whose name begins with a B or by all physical states instantiating a neurological state that manifests belief in dualism. That aspects pick out physical types, not just tokens, is crucial for this understanding of the term 'aspect' and crucial for radical physicalism and radical reductionism. Let us call this option the HS option, for Hemmo and Shenker, or the *flat* option as it is called by them.[1] I will later examine this notion of flatness and inquire in what sense the HS option is indeed more flat than its D alternative, but already at this point we are getting a sense of the importance of the notion of description and the role it plays in the two options I have distinguished. Acknowledging this role clarifies the notion of level (and stratification into levels) underlying the two options.

At some point, but here I am less sure about the details, the HS option was augmented by the *observer* (Shenker, 2015).The idea is—very schematically—that the physical type that is shared by dualists must include the physical brain states of those who perceive them as dualists. Again, this modification will be taken up at a later stage. For now, we should note that some progress has been made, for in distinguishing the two versions of aspect-language, we are no longer talking merely about names, but about the meaning and substance of the physicalist position. Orly, Meir, and I have been debating this issue back and forth; I thank them for the stimulation and pleasure our conversations have given me.

[1] I heard the term 'flat physicalism' for the first time in the 2019 conference on which this volume is based, but the position is one that Shenker and Hemmo have been defending for at least a decade. For their recent writings on the subject see their "Why Functionalism is a Form of 'Token Dualism'" in this volume, their "A Dilemma for Davidson's Anomalous Monism" http://philsci-archive.pitt.edu/19563, and their (forthcoming) "Flat Physicalism".

Davidson's anomalous monism has been taken on by Hemmo, Shenker, and other critics such as Jaegwon Kim, all of whom see it as a form of dualism. They sometimes concede that it is property-dualism rather than substance-dualism and that it therefore differs from Descartes' dualism, but this is a minor difference in their view. If, from the very start, dualism is supposed to be wrong—a deviation from the right path of pure physicalism—then the mere fact that a position amounts to dualism should dissuade us from accepting it. But it would be preferable to *show* that dualism is wrong rather than presuppose it, that is, it would be better to demonstrate that any position that amounts to dualism has some built in fault that makes it unacceptable. With regard to nonreductive physicalism, then, the question is not solely whether it commits one to dualism, but whether it involves some more serious problem such as incoherence, conflict with our best scientific theory, and so on. Worries of this sort have indeed been expressed, the most serious among them being that dualism clashes with the causal closure of physics. In what follows I revisit some of these arguments and try to show where they go wrong. Thus, I critique flat physicalism and its assertion that higher level descriptions are always partial descriptions (aspects) of lower-level entities (states, events). I thereby defend nonreductive physicalism, multiple realizability, and the possibility of token identity without type identity.[2] But before doing so it would be useful to go over some familiar ground and recap the standard accounts of reductive and nonreductive physicalism. Here is what follows. In Sect. 2.2 I review the notions of law (in particular in contrast with that of an accidental generalization), reduction, and nonreductive physicalism. Section 2.3 focusses on the concept of entropy as a test case for flat physicalism and reductionism. Kim's critique of Davidson is the subject of Sect. 2.4. I conclude, in Sect. 2.5, with a variation on Davidson's position and a note on the hierarchy of physical levels.[3]

2.2 Laws, Reduction and Nonreductive Physicalism

(a) Laws and accidental generalizations: To begin with, recall that laws (regardless of whether they are formulated in mathematical, physical, or everyday language) refer essentially to *types*. Even when they are applied to individual entities, processes, states and events, they refer to them *under a description*, namely, they involve types.[4] In order to apply a law to a particular situation

[2] I defend Davidson's version of nonreductive physicalism, which is the version that comes under Kim's attack. I believe, however, that my defense applies, *mutatis mutandis*, to other versions.

[3] In what follows I draw on my book on causation (Ben-Menahem, 2018), but my focus here is different, resulting in a different organization and of the arguments.

[4] This point is implied by Hempel's account of explanation and stressed by Davidson in "Causal relations" (1980a), where he distinguishes causal relations, which are not sensitive to the description of the related events, from explanatory statements, which are. See also Steiner (1983), who credits Sydney Morgenbesser with the same insight.

it is therefore necessary to describe the situation properly, that is, describe it in terms of predicates that match those appearing in the law. Many other descriptions may be correctly applied to the same situation (event, process and so on), but they will not make it subsumable under the said law. In addition to their familiar roles in predicting and explaining phenomena, I should stress that laws also serve to *characterize* the types of entities, states, and events that fall under them. Pauli's principle characterizes fermions and the speed (in vacuum) of 299.792458 km. per second characterizes electromagnetic radiation, As Goodman has convincingly argued in *Fact, Fiction, and Forecast* (1955), there is a close relationship between laws and the types they invoke. Laws are projectable and so are the predicates that designate the law-covered types. I don't need to break this particular vase in order to find out whether it is fragile. I know it is made of glass; I know that glass (in room temperature) is fragile. Because laws and the types they characterize are inseparable, I can project frangibility from one piece of glass to another, but not from glass to other types such as plastic.

Typically, accidental generalizations also refer to types, but neither the generalizations, nor the predicates describing the types they invoke are projectable. It may happen that all the vases in my house on 1.1.2021 are made of glass, but no information can be gleaned from this generalization regarding vases in general, vases in other homes, vases in my house on 1.1.22 and so on. Goodman noted further that whereas laws are confirmed by their instances, but not by any particular instance, accidental generalizations, can only be confirmed by checking each one of their instances. Although accidental generalizations also pick out aspects of the individual entities (states, events) they apply to—being in my house picks out an aspect of the vases that happen to be there—from the scientific point of view these aspects are inert; they do not enhance our knowledge beyond what we had already established. The type of persons whose family name begins with a B is perhaps somewhat more projectable; we can predict that names of members in this set would appear in the telephone directory above those of C persons, or that their sons will probably also belong to the same set, and so on. But these projections depend on human conventions; fundamental physics, it seems, is blind to the set of B persons and cannot provide a physical criterion that singles out (the physical state of) B persons from all other physical entities.

(b) Reduction: According to Ernst Nagel's classic account (1961, Chapter 11), reduction requires that the concepts of the reduced, higher-level, theory be defined in terms of concepts belonging to the theory of the fundamental-level, and that the higher-level laws of the reduced theory be derived from the laws (of the theory) of the fundamental level.[5] In view of the paucity of

[5] This formulation may not be faithful to the letter of Nagel's account, but is consonant with its sprit. Note that I am only discussing what Nagel (1961, p. 342) refers to as "heterogeneous reduction." As Nickles (1973) observed, there is an opposite usage of the notion of reduction,

examples that satisfy these strong requirements, they are often weakened in the following way. The definitions in question need not establish the synonymy of the defined (reduced) terms with the defining terms, but rather, the definitions can be empirical laws (bridge laws) establishing co-extensionality rather than synonymy. And the laws derived from fundamental-level laws need not be identical to the laws of the reduced (higher-level) theory; it suffices that they constitute good-enough approximations of these higher-level laws. The fundamental laws can, for example, yield a probabilistic version of the laws of the reduced theory, as in the derivation of thermodynamics from statistical mechanics.[6]

Given that critique of nonreductive approaches is usually couched in terms of causation (e.g. threats to the causal closure of physics), we may benefit from a reformulation of the foregoing account of reduction in causal terms. Reduction then requires that for each higher-level causal relation or process we can point to an underlying causal relation or processes taking place at the fundamental level of physics. When reduction of this kind is achieved, genuine causation exists only at the fundamental level. As there is no consensus on the meaning of causation, the causal criterion for reduction is more ambiguous than the Nagelian. For instance, depending on whether or not we understand causation in terms of lawful regularities, the two formulations can be seen as competing or complementary. In any event, on both versions, successful reduction makes the reduced higher-level theory redundant. It is redundant from the explanatory point of view because the laws of the fundamental theory provide all the explanations provided by the reduced (higher-level) theory and it is likewise redundant from the causal perspective because the causal network associated with the reduced theory is replaced with that of the fundamental level.

(c) Nonreductive physicalism. Accepting either one of the above characterizations of reduction does not commit one to the belief that *all* higher-level theories are reducible in this way. It is this additional commitment to overall reducibility—reduction*ism*—that distinguishes flat physicalism from nonreductive physicalism. Reductionists assert that all the concepts of higher level theories are definable in terms of (or are at least co-extensional with) concepts of fundamental theories. Since concepts correspond to types, it follows that (according to reductionism) all the types characterized by higher level laws correspond to types characterized by fundamental laws. But flat physicalists like Hemmo and Shenker (if I understand them correctly) actually assert

common among physicists, on which it is the fundamental theory that is reduced to the higher-level theory, meaning that the former converges on the latter in the limit. Thus, one might say that special relativity reduces to Newtonian mechanics at velocities much lower than that of light ($v \ll c$). I will use 'reduction' in the philosophers' sense, which is more apt for discussing the problems that concern us here.

[6] In this passage I ignore the current debate on the success of the reduction of thermodynamics to statistical mechanics; See Hemmo and Shenker 2012 and the literature they cite. The reducibility of the concept of entropy is discussed in section III.

> much more: they maintain not only that concepts and types of higher-level *theories* correspond to fundamental physical types, but that higher-level types *of any kind*, whether or not they figure in some scientific theory, correspond to fundamental physical types. Thus, as far as reducibility is concerned, concepts such as 'dualist' and 'B-person' are (in their view) on a par with the concepts of pressure and temperature even if only the latter have so far been successfully reduced.

Generally, nonreductive physicalists do not contest reduction within science; it is the correspondence of non-scientific types to physical types that they deny. The controversy on nonreductive physicalism has focused on mental events, which according to Davidson (1980b) are physical events that have mental descriptions. (The question of whether mental events are unique in exemplifying nonreductive physicalism will be addressed later) Like other descriptions, mental descriptions assemble individual events into types, but these types, Davidson contends, are not 'held together 'by laws, neither mental laws nor physical ones. And though every mental event is a physical event, the physical events that instantiate a particular mental type do not constitute a physical type and are therefore not characterized by physical laws in the way that Pauli's principle singles out fermions. Reduction, as characterized above is therefore blocked. Mental states supervene on physical states, but due to multiple realization, they do not create reducible types. And since, where there are no types, there are no laws, the argument leads to the anomalous nature of the mental. Davidson made his position seem paradoxical by showing that it enables him to combine a number of seemingly contradictory claims: every mental event is a physical event; there are causal relations between the mental and the physical; causal relations entail the existence of laws; there are no mental laws.[7] But in the light of the forgoing discussion of the different options for understanding physicalism and aspect-language, Davidson's solution is not as paradoxical as it at first seems. Still, his solution has met with serious objections that I discuss in Sect. 2.3. I begin, however with an example of reduction in science.

2.3 Reducibility and Multiple Realization in Statistical Mechanics

When surveying the literature on the success (or failure) of the reduction of thermodynamics to statistical mechanics, one usually finds that it centers on irreversibility. The problem generated by irreversibility is that it seems impossible to derive the *a*symmetry built into the second law of thermodynamics from the

[7] Davidson's commitment to the Humean position that there is no causality without regularity is not actually essential for the main points of either "Causal Relations" (1980a) or "Mental events" (1980b) but it certainly adds to the magical appearance of his solution.

underlying time-symmetric laws of mechanics.[8] This problem pertains to Nagel's second requirement—the derivation of the *laws* of the higher-level theory from those of the fundamental one. I seek, however, to address Nagel's first requirement, the reduction of thermodynamic *concepts*, which, though conceptually prior to that of deriving the second law is mostly neglected.

In principle, we can have different descriptions of a thermodynamic system. In particular, we can entertain a micro-description specifying the values of each one of its physical parameters for every one of its constituent particles, and a macro-description in terms of its macro-observables, such as its pressure, volume and temperature. As it happens, the former description is unavailable to us, the macro-creatures that we are, while the second is easily obtained. Whereas classical thermodynamics was formulated in terms of macro-descriptions alone, statistical mechanics seeks to connect the two levels of description. The realization that such a connection exists was driven by the kinetic theory of heat on which heat is an expression of the incessant movement of huge numbers of particles, moving and interacting according to the laws of classical mechanics. For some macroscopic parameters, the connection with micro-properties is relatively clear—it is quite intuitive, for instance, to correlate the pressure exerted by a gas on its container with the average impact (per area unit) of micro-particles on the container. But this is not the case of other macro-properties, entropy, in particular. Recovering the notion of entropy is essential for the recovery of the second law of thermodynamics and constitutes a major objective of statistical mechanics.

The fundamental insight underlying the connection between entropy and the micro-level is the following: Macrostates are multiply realizable by microstates, that is, the same macrostate could be realized by numerous different microstates. The implication is that in general, the detailed description of a system's microstate plays no role in the macro-description of the system and its evolution. The temperature of a macrostate, for example, is defined as the average velocity of the molecules comprising the underlying microstates and clearly, there are multiple ways of getting the same average. Despite this multiplicity, though, the higher-level concept of temperature still reflects a property exhibited at the fundamental level. What about entropy? By contrast with temperature and pressure, there is no property, or aspect, or partial description of the microstate of a system that is picked out by, or corresponds to, its entropy. What matters for the definition of entropy is only the *number of ways* (or its measure-theoretic analogue for continuous variables) in which a macrostate could be realized.[9] As long as we have access to these numbers (or their measure-theoretic analogues) and can use them to distinguish between

[8] The asymmetry is manifest in the second law's proclaiming that (very roughly), in an isolated system, entropy can spontaneously increase but not decrease)

[9] This special character of entropy, as a result of which it is not directly measurable, has led to the extreme position that it is not a physical quantity. I do not share this position but cannot argue this point here. Present day writers emphasize that using the Lebesgue measure for probability (or for entropy) in this context is not the only possibility and is therefore a non-trivial, albeit intuitive, assumption. See, for example, Hemmo and Shenker (2012b), Pitowsky (2012).

different macrostates, the fact that the detailed description of the actual microstate remains hidden is no obstacle. Entropy thus reflects a property of macrostates, not a property of the underlying level of molecules. To put it in more picturesque terms, if molecules had consciousness, they could perhaps be interested in their average velocity and thus (even though they had no concept of macrostate) have a concept that corresponds to the macro-concept of temperature. But without the notion of a macrostate they could not even entertain the multitude of a macrostate's possible realizations and would altogether miss the notion of entropy.[10]

There are a number of lessons to draw from the case of entropy. First, higher-level concepts are essential and cannot be eliminated in favor of lower-level ones. As we have seen, entropy does not reside entirely at the fundamental level; it requires the concept of macrostate and its 'size' to be defined. The HS version of construing descriptions as referring to aspects of the fundamental physical state thus comes under pressure. Moreover, macrostates and their 'size' play an essential explanatory role. If, for instance, we ponder the stability of one particular macrostate—the equilibrium— relative to other macrostates, we need to refer to the entropy or probability of the macrostates in question and these magnitudes, we have just seen, are not completely reducible to the fundamental level. Second, description-sensitivity is salient: macrostates do not descend from heaven with fixed identities—they are given an identity by the description we use. This does not mean macrostates are fictions—they are as real as microstates—but it means that in order to understand their behavior *as macrostates*, that is, to discover the laws that govern their behavior as macrostates, they must be characterized in a useful way. This characterization is given by us and goes beyond micro-properties of the system. Third, the correlation between microstates and macrostates is a many-one relation (or function) that manifests supervenience in the sense given to the term by Davidson. If a microstate is specified, the question of whether it realizes a certain macrostate receives a determinate answer, but the converse does not follow; the identification of a macrostate does not determine which microstate realizes it on this particular occasion.[11] The many-one function is also the formal representation of the *insensitivity* of the scope of the function to many of the features that distinguish its arguments from one another. The important characteristic of a macrostate—

[10] The standard formalism that captures this relation between microstates and macrostates is the representation of the former by points, and the latter by regions, in the 6 N dimensional phase space (where a point represents a microstate of the entire system in terms of 6 co-ordinates for each one of its N constituent particles, e.g. 3 co-ordinates for position and 3 for momentum). The idea, then, is that each macrostate is realizable by all the microstates corresponding to points that belong to the volume representing this macrostate—clearly a volume that can vary enormously from one macrostate to another. This insight led to the identification of the volume representing a macrostate in phase space with the probability of this macrostate and to the definition of entropy in terms of this probability. (This ahistorical account is closer to Boltzman than to Gibbs.) As the number of points is infinite one actually needs to talk of a measure rather than simply of numbers. See note 9 above for references.

[11] Entropy supervenes on the microstate of the system in Botzmamm's statistical mechanics but it is not clear whether the same holds for entropy in the Gibbs formulation of statistical mechanics.

its probability—is relatively indifferent to many of the details characterizing the microstates that belong to this macrostate.

Insensitivity of the evolution taking place on a higher level to the detailed structure of the fundamental level is not unique to the case of entropy. Another example is provided by the phenomenon known as universality: the strikingly similar behavior of very different physical systems at (or close to) specific points—*critical* points. (Batterman, 2002; Morrisson, 2012, 2015). Water and ferromagnetic materials have little in common in terms of their physical/chemical structure and behavior. But during phase transitions such as the water's freezing and the ferromagnet's magnetization, unexpected similarity appears not only in the overall pattern of symmetry-breaking that these transitions involve, but also in the precise values of parameters—critical exponents-—that determine the characteristics of these transitions. The mechanisms are clearly distinct; electron spins, for example, play a crucial role in magnetization, but not in freezing or condensation. But the similarity between the systems manifesting universality reveals the overall pattern's insensitivity to structural and dynamic details at the fundamental level.[12] Although it is sometimes questioned whether the theory that explains universality is a physical theory, or merely a mathematical technique,[13] the situation with regard to reduction is quite similar to that of reduction in statistical mechanics. Every system exhibiting universality satisfies the requirement that higher-level patterns supervene on underlying micro-structures, but the overall patterns and the parameters that characterize them are not derived solely from the fundamental laws.

The examination of the reduction of thermodynamics to statistical mechanics demonstrates that salient features of nonreductive physicalism, multiple realizability, supervenience, sensitivity to description, and insensitivity of higher levels to specifics of the lower ones, are part and parcel of physics.

2.4 Meeting Kim's Objections to Nonreductive Physicalism

The complaint that a certain position leads to dualism, as noted in the introduction, is not sufficient to undermine this position. To be convincing, some more serious argument against the position in question should be adduced. With regard to

[12] This insensitivity is thought to reflect the fact that at (or near) critical points there is a change in the nature of the coupling between components of the system and the range of their relevant interactions. Whereas under normal conditions long-distance coupling and correlations can be ignored, at critical points this idealization is no longer valid and all interactions must be taken into account. Calculation of these overwhelmingly complex processes is made possible by the technique known as the renormalization group, which involves iterative coarse-graining of the system, with the result that the behavior of the system on every coarse-grained level is analogous to the behavior manifested on the preceding (more fine-grained) level. In the course of this iterative process, the differences between levels within the same system, and the differences between the dynamics of different systems, are washed out.

[13] See Morrison (2012) and the references cited there.

nonreductive physicalism, the major concern pronounced by its opponents is its sanctioning *downward causation*, thereby allowing for violation of the *causal closure* of the fundamental level of physics. Taking Jaegwon Kim as a representative of this line of critique, I address his concern and argue that it is unfounded.

To get a feeling for the possible relations that may obtain between different levels, consider a fundamental level (or theory) F and a higher level (or theory) H. At the outset, we should note that the laws of F and the laws of H could be consistent or inconsistent with each other. As already mentioned, there are actually very few cases where higher-level theories are rigorously consistent with lower-level ones; typically, the laws of the basic level contradict those of the higher level.[14] But we can agree to settle for a weaker condition than perfect consistency—one theory can, for instance, be consistent with a good-enough approximation of the other—and assume that this condition is satisfied in the case of F and H. There are still at least three possibilities:

1. Reduction: All H-laws can be reduced to F-laws, so that H-laws are eliminated in favor of F-laws. In this case H-laws are redundant, and phenomena on H are deemed epiphenomena.
2. Lacunae: There are H-laws that cover (predict and explain) phenomena that F-laws do not cover.
3. Overdetermination: There are H-laws that are irreducible to F-laws, but provide alternative predictions and explanations of phenomena that F-laws suffice to explain. Being entailed by two distinct sets of laws, these phenomena are thus overdetermined.

Similar relations can be formulated in terms of causality:

1. Reduction: All H-causes are actually F-causes, rendering H-causes redundant.
2. Lacunae: Some H-causes bring about effects that have no F-cause.
3. Overdetermination: Some H-causes, though irreducible to F-causes (that is, though not identical to any F-cause), bring about effects that F-causes also suffice to bring about. These effects are therefore overdetermined.

Denying the possibility of lacunae and overdetermination, reductionists see only the first option as viable. Their reasoning involves the deterministic assumption of the physical closure of the basic level: the assumption that every basic-level event is determined (explicable, predictable) by the laws and initial conditions (or boundary conditions) of that level. This deterministic assumption only holds for closed systems and is valid only for classical theories, not quantum mechanics. Nevertheless, if, for argument's sake, the assumption of physical closure is accepted,

[14] This is clearly the situation in statistical mechanics—the reductionists' favorite paradigm case–but it is also what happens in simpler cases that are usually thought of in terms of generalization rather than reduction. Strictly speaking, Newtonian mechanics contradicts Galileo's law of free fall, but the affinity between the two theories' respective predictions for small enough terrestrial distances induces us to think of Galileo's law as an instance of Newton's more general law.

lacunae and overdetermination are ruled out. Reductionism is vindicated, or so it seems.

According to Kim, the physical closure is threatened by nonreductive physicalism: "The physical causal closure remains very much a problem within the stratified ontology of non-reductivism. Nonreductive physicalism, like Cartesianism, founders on the rocks of mental causation" (1993) p. 339. Furthermore, Kim construes his opponent as claiming that "mentality . . . takes on a causal life of its own and begins to *exercise causal influence "downward" to affect what goes on in the underlying physical-biological processes".* (Ibid p. 349; italics in the original). The danger, then, is downward causation, which is implied, according to Kim, by nonreductive physicalism. Let us take a closer look at his argument. Supposes, with Kim, that a higher-level property M is causally efficacious with respect to another higher-level property M* and suppose further that these higher-level properties are instantiated by fundamental properties P and P*.

Then, according to Kim,

> We seem to have two distinct and independent answers to the question "Why is this instance of M* present?" *Ex hypothesi*, it is there because an instance of M caused it; that's why it's there. But there is another answer: it's there because P* physically realizes M* and P* is instantiated on this occasion. I believe these two stories about the presence of M* on this occasion create a tension. (Ibid p. 351)

He continues

> Is it plausible to suppose that the *joint* presence of M and P* is responsible for the instantiation of M*? No; because this contradicts the claim that M* is physically realized by P*. . . . This claim implies that P* alone is sufficient to bring about M*. . . . And the supposition is also inconsistent with our initial assumption that a given instance of M was a sufficient condition for that instance of M*. (Ibid p. 352)

Here, obviously, Kim conflates the relation of instantiation which is in any specific case an *identity*, with that of a causal relation between two different events. P* in the case he considers does not bring about M*— it is identical with an instance of M*. The cause of P* (on the assumption he has made) is another micro-event P and the relation between them may be lawful and deterministic. So there is no overdetermination. It is possible, as we saw, and consistent with fundamental physics, that the causal relation between P and P* will not provide a full explanation

of the relation between M and M*. The relation between two macrostates in statistical mechanics may depend on their relative stability and thus on their 'size,' which is not reducible to a property of any single microstate This should not indicate any explanatory lacunae on the basic level, but merely a change of our explanandum, which is now a different type of event or state (one that comprises numerous microstates) and therefore calls for a different explanation.

It appears that Kim is misled by the upwards downwards metaphor, saddling his opponent with a picture of the higher level as inhabited by upstairs folk, who intrude on their downstairs neighbors and prevent them from going on their business. This is an abuse of the metaphor and a misunderstanding of Davidson.[15] There are no two sets of neighbors. Higher levels, as levels of description, are linked to the lower levels by various kinds of identities, not by causal connections that could interfere with the causal network of the lower level. Ironically, in this case it is Kim who slips back into dualism.

Kim has another argument against the causal efficacy of upper-level properties. It is based on a principle which he calls "The Causal Inheritance Principle":

> If mental property M is realized in a system at t in virtue of physical realization base P, the causal powers of *this instance of M* are identical with the causal Powers of P. (1993 p. 326; italics in the original)

In one sense the principle is trivial. The causal powers of this instance of M are indeed the causal powers of the physical state that realizes it, but this is true simply because this instance of M *is* a P state, so that there is only one entity exerting whatever causal influence it has. The idiom of inheritance, though, is misleading, suggesting two distinct entities one of which inherits something from the other. Could there be a less trivial sense of the principle, for instance, the principle that the causal powers of M are inherited by every one of its realizers? But on this reading the principle is wrong. In statistical mechanics, we saw, the causal efficacy of macrostates qua macrostates (and their explanatory import) is not inherited by every microstate that realizes them. My conclusion is that, Kim's arguments notwithstanding, nonreductive physicalism is perfectly consistent with the physical closure of the fundamental level.

2.5 Concluding Remarks

My model of nonreductive physicalism was Davidson's anomalous monism which focusses on mental events. The mental is also at the heart of Kim's arguments critiqued in the previous section. Because of the long history of the debate on the mind-body relationship, the focus on the mental is understandable, but it tends to blind us to other applications of Davidson's insights. In Sect. 2.3 I compared

[15] Davidson would also object to Kim's talk of causes as sufficient conditions, but let's not be pedantic.

Davidson's account of the relation between the mental and the physical with the relation between macrostates and microstates in statistical mechanics. The analogy is doubly encouraging for nonreductive physicalism. First, if multiply realized (higher-level) types such as entropy are indispensable in statistical mechanics, it alleviates the worry that multiple realization in inconsistent with physics. Second, it suggests that the mental is not unique in exhibiting a failure of reducibility. Indeed, one need not confine the discussion to the mental (or stretch the imagination as far as Goodman's grue-some predicates), to find examples of concepts (predicates) that defy law-likeness and projectability. I mentioned some examples in the introduction, B-named persons, say, but there are also very simple physical objects whose description is irreducible to their physical characteristics. Consider a stop sign. It is certainly a physical object and belongs to the category of physical objects. It obeys the laws of physics (depending on its makeup it may obey different laws), and does not threaten to overrule any of the laws or causal relations on the fundamental level. Nonetheless, there is no physical category that corresponds to the category of stop signs. Not just because stop signs are multiply realizable (which, of course, they are), but because the concept of stop sign is open ended and non-projectable.[16] Any number of objects could become stop-signs and no physical property, or structure, or set of specific laws, distinguishes stop signs from other objects. The description 'stop sign' thus refers to an aspect of the physical state of objects falling under that description, but in this case aspect language must be understood in accordance with the D option outlined in the introduction, not the flat, HS option. Examples of this kind suggest amending Davidson's point about the mental. It is not a dramatic amendment, for after all, stop signs are symbols, requiring an interpreting mind to understand them. Their open-endedness thus derives from their symbolic significance and is ultimately predicated on mental activity. Such examples do suggest, however, that the crucial feature differentiating the lawful from the lawless in this context is symbolic meaning rather than mentality *per se*. Even if mental states of fear, surprise, and so on, were discovered to correspond to neurological types (which is perhaps not unreasonable) or physical types (which is far less likely), the property of being frighten*ing* or surpris*ing* would still be open-ended and lawless. Under the description of being frightening or surprising, the events and entities falling under these descriptions would be open-ended and lawless as well.

If symbolic meaning is what counts in the above examples, the idea of including the observer, or interpreter, in the reduction does indeed suggest itself. Thus, one could salvage the HS version by arguing that the category of stop signs does not correspond to a physical type that all and only stop signs instantiate, but to a physical type that includes in addition all and only physical states of brains that perceive the objects at hand as stop signs. We are back to square one, however, for the Davidsonian maintains that there is no reason to think that this type of brain state

[16] Multiple realizability in itself does not entail open-endedness. Universality, as we saw, is linked to multiple realizability, but it is conceivable that it is only exhibited in a specific kinds of systems and is not open-ended in the way that the concept of stop sign is.

corresponds to a physical type whereas the HS physicalist insists that it does. At this point there is no decisive argument on either side (accept for the charge of dualism directed at Davidsonians by their opponents, but I said I wouldn't argue about names). Even at this impasse, however, we should note that reference to the observer opens a wide gap between ordinary physical types and observer-including types. Fermions and electromagnetic radiation are characterized by physicists without the mind that perceives them as such and if stop signs cannot be so described, there is a significant difference between the physical nature of the former and (the alleged) physicality of the latter.

Finally, let us revisit the notion of level that is at play in the forgoing discussion of reducibility. What nonreductive physicalists take to be irreducible are types characterized by certain descriptions. In the examples considered in this paper there was no commitment to different levels of reality, only to different ways of grouping the elements of reality (whether they are states, events, or objects) into sets that fall under various descriptions. (I am not saying that stratification into levels of reality is impossible, it may well be useful, only that such stratification was not at issue in this paper). Some of these sets, we saw, correspond to physical types and when they do their higher-level descriptions pick out the very same sets that are picked out by their descriptions in the language of fundamental physics. Here the higher-level language renames or redescribes physical types that have proved to be projectable. We thereby obtained a translation of the higher-level language into the fundamental one; the first step towards reduction of the relevant higher-level theory to the fundamental one has been made. For other sets, illustrated by the example of entropy, the correspondence between higher-level and lower-level types is far more complex. Nonetheless, entropy does not confront us with a new level of reality, but rather with a new level of description. The pressing question about the feasibility of reduction, I submit, is not whether a multileveled structure could be collapsed (without damage) to its basis, but whether the various descriptions of reality could be translated in their entirety into one particular privileged description. Reductionists and nonreductionists could agree on this formulation of the problem. If they do, then despite their disagreement about the reducibility of each and every description, their corresponding visions of reality could, it seems, be equally 'flat.'

References

Batterman, R. W. (2002). *The devil in the details*. Oxford University Press.

Ben-Menahem, Y. (2018). *Causation in science*. Princeton University Press.

Davidson, D. (1980a [1967]). Causal relations. In *Essays on actions and events* (pp. 149–162). Clarendon Press.

Davidson, D. (1980b [1970]). Mental events. In *Essays on actions and events* (pp. 207–224). Clarendon Press.

Goodman, N. (1955). *Fact, fiction, and forecast*. Harvard University Press.

Hemmo, M., & Shenker, O. (2012a). *The road to Maxwell's demon*. Cambridge University Press.

Hemmo, M., & Shenker, O. (2012b). Measures over initial conditions. In Y. Ben-Menahem & M. Hemmo (Eds.), *Probability in physics* (pp. 87–98). Springer.
Hemmo, M., & Shenker O. *A dilemma for Davidson's anomalous monism*. http://philsci-archive.pitt.edu/19563.
Hemmo, M., & Shenker, O. (forthcoming). Flat physicalism. *Theoria*.
Kim, J. (1993). *Supervenience and mind*. Cambridge University Press.
Morrison, M. (2012). Beyond calculation: Extracting physical information from mathematical methods. *Iyyun, 61*, 149–166.
Morrison, M. (2015). *Reconstructing reality: Models, mathematics, and simulations*. Oxford University Press.
Nagel, E. (1961). *The structure of science*. Routledge and Kegan Paul.
Nickles, T. (1973). Two concepts of intertheoretic reduction. *Journal of Philosophy, 70*, 181–201.
Pitowsky, I. (2012). Typicality and the role of the Lebesgue measure in statistical mechanics. In Y. Ben-Menahem & M. Hemmo (Eds.), *Probability in Physics* (pp. 41–58). Springer.
Shenker, O. (2015). (Hebrew) Davidson on descriptions: A principle theory. *Iyyun, 64*, 171–190.
Steiner, M. (1983). Under a description. How many questions? In L. S. Cauman, I. Levi, C. Parsons, & R. Schwartz (Eds.), *Essays in honor of Sidney Morgenbesser* (pp. 120–131). Hackett.

Chapter 3
The Quantum Field Theory on Which the Everyday World Supervenes

Sean M. Carroll

Abstract Effective Field Theory (EFT) is the successful paradigm underlying modern theoretical physics, including the "Core Theory" of the Standard Model of particle physics plus Einstein's general relativity. I will argue that EFT grants us a unique insight: each EFT model comes with a built-in specification of its domain of applicability. Hence, once a model is tested within some domain (of energies and interaction strengths), we can be confident that it will continue to be accurate within that domain. Currently, the Core Theory has been tested in regimes that include all of the energy scales relevant to the physics of everyday life (biology, chemistry, technology, etc.). Therefore, we have reason to be confident that the laws of physics underlying the phenomena of everyday life are completely known.

3.1 Introduction

Objects in our everyday world—people, planets, puppies—are made up of atoms and molecules. Atoms and molecules, in turn, are made of elementary particles, interacting via a set of fundamental forces. And these particles and forces are accurately described by the principles of quantum field theory.

We don't know whether relativistic quantum field theory is the right framework for a complete description of nature, and indeed there are indications (especially from black hole information and other aspects of quantum gravity) that it might not be. But if we imagine describing nature in terms of multiple levels of reality, one such level appears to be a particular kind of quantum field theory, with other levels above (e.g. atoms and molecules; people and planets and puppies) and possibly other levels below.

S. M. Carroll (✉)
Walter Burke Institute for Theoretical Physics, California Institute of Technology, Pasadena, CA, USA

Santa Fe Institute, Santa Fe, NM, USA

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_3

In addition to a "vertical" division into levels, we can also consider carving each level "horizontally" into different regimes, corresponding to different kinds of physical situations. We might, for example, have a pretty good idea of how certain human beings will behave under ordinary conditions, but be less confident in how they will behave in extreme circumstances. Within the domain of physics, we might distinguish between different regimes of energy or temperature or physical size.

In this paper I focus on the level of reality described by quantum field theory, in what we might call the "everyday-life regime" (ELR)—the energies, densities, temperatures, and other quantities characterizing phenomena that a typical human will experience in their normal lives. This doesn't just mean, for example, the kinetic energy per particle that a human can muster under the power of their own musculature; it also includes phenomena such as sunlight that ultimately involve more extreme conditions in order to be explained. It does not include conditions in the early universe, or near neutron stars or black holes, or involve phenomena such as dark matter and dark energy that don't interact noticeably with human beings under ordinary circumstances.

Modern physics has constructed an "effective" quantum field theory that purports to account for phenomena within this regime, a model that has been dubbed the "Core Theory" (Wilczek, 2015). It includes the Standard Model of Particle Physics, but also gravitation as described by general relativity in the weak-field limit. I will argue that we have good reason to believe that this model is both *accurate* and *complete* within the everyday-life regime; in other words, that the laws of physics underlying everyday life are, at one level of description, completely known. This is not to claim that physics is nearly finished and that we are close to obtaining a Theory of Everything, but just that one particular level in one limited regime is now understood. We will undoubtedly discover new particles and new forces, and perhaps even phenomena that are completely outside the domain of applicability of quantum field theory; but these will not require modifications of the Core Theory within the ELR, nor will the Core Theory fail to account for higher-level phenomena in that regime. (A nontechnical version of this argument was given in Carroll (2017).)

The interesting part of this claim is that it relies specifically on features of quantum field theory, which distinguish this paradigm from earlier models of physics. In particular, the effective field theory paradigm gives us good reason to believe that the dynamics of the known fields are completely understood, and the phenomenon known as "crossing symmetry" implies that any new particles or forces must interact too weakly with Core Theory fields to be relevant to everyday-life phenomena. In this paper I will explore this claim, starting with a precise statement of what the argument is supposed to be, and then a summary of the effective-field-theory approach. I then discuss the specifics of the Core Theory, including why we are confident that its dynamics are understood in the ELR. Then we will move to the feature of particle physics known as crossing symmetry, and how it constrains

the possibility of unknown fields. I will then discuss the implications of these ideas for physics more broadly, and the wider project of understanding levels of reality.

3.2 What Is Being Claimed

The structure we are considering is portrayed in Fig. 3.1, with levels of reality arranged vertically. The middle ellipse is an effective relativistic quantum field theory, including weak-field quantum general relativity, thought of as a field theory on a flat background spacetime. The smaller ellipse is the Core Theory of known particles and forces, with additional unknown particles and forces in the rest of the region. The top ellipse summarizes all the more macroscopic levels, and is divided into the everyday-life regime (ELR) in the small ellipse, and more extreme astrophysical phenomena elsewhere. (For our purposes here we can classify things like ultra-high-energy cosmic rays as astrophysical.) Finally, we include a hypothetical level below, and therefore more fundamental than, effective quantum field theory. I will refer to the theoretical explanations for what is described by each box as "theories" or "descriptions" or "models," interchangeably.

The arrows in this figure indicate what phenomena depend on what other sets of phenomena; solid arrows are known relations, and dashed arrows are plausible but unknown. The important claim being made is that certain arrows one could imagine drawing—from "Everyday life" to "Unknown particles and forces" or "Underlying reality"—do *not* appear. In particular, everyday macro phenomena do not depend on either new particles/forces, nor directly on the underlying reality. The Core Theory

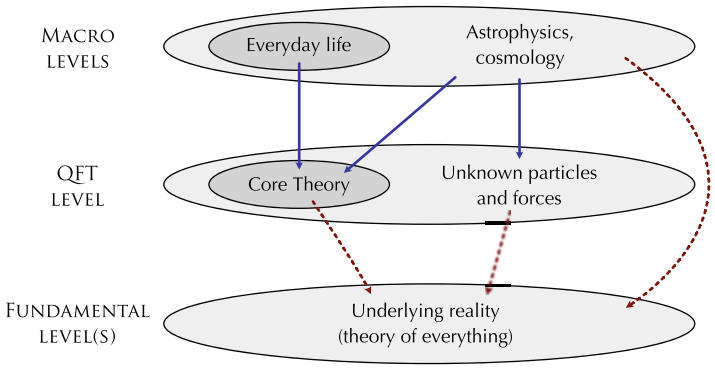

Fig. 3.1 Direct dependency relations between sets of phenomena at different levels. Solid blue arrows are established, while dashed red arrows are conjectural. Arrows that could be drawn, but are not, are relations we have good reason to think do not exist. So phenomena in the everyday life regime depend on the Core Theory, but not on unknown particles and forces, nor (directly) on an underlying theory of everything. Astrophysical phenomena depend on both the Core Theory and on new fields, and may depend directly on the underlying theory (e.g. in regimes where quantum gravity is important)

provides a complete and accurate description, we have good reason to believe, of everything on which macroscopic phenomena in the ELR supervene. (In the next section we will be more specific about what is meant by the ELR.)

To make this claim more precise, let us distinguish between the Core Theory, which we know, and the idea of the Laws of Physics Underlying Everyday Life (LPUEL), whatever they might actually be. We take it as established that everyday objects are at least partly made up of atoms, which are at least partly made of elementary particles, and that in some circumstances these particles interact through fundamental forces according to the standard understanding of physics, at least approximately. The LPUEL, then, is whatever set of ingredients and dynamical rules operating at what we usually think of as the level of elementary particles that suffices to account for the properties of phenomena we experience in everyday life. The Core Theory is a specific model, which we are arguing completely captures the LPUEL. In principle, we might imagine a wide variety of ways in which the LPUEL deviate from the Core Theory; there might be heretofore undiscovered particles or forces that are relevant to the behavior of macroscopic phenomena, or quantum field theory itself might break down even within the ELR. Our claim is that we have good reasons to believe this doesn't happen.

The argument will be as follows:

1. We have good reasons to believe that the LPUEL take the form of an effective quantum field theory (EQFT).
2. The Core Theory is an EQFT that to date is compatible with all known experimental data within the everyday-life regime.
3. Within the EQFT paradigm, the Core Theory could be modified in two possible ways: we could modify the dynamics of the known fields, or introduce additional fields.
4. Modified dynamics that could affect the LPUEL would require gross violations of the expectations of the EQFT paradigm, and are constrained experimentally.
5. Experimental constraints also imply that additional fields would be either too massive, too weakly-coupled, or too rare to affect the LPUEL.
6. Therefore, we have good reason to believe that the LPUEL are completely known.

It's worth being especially careful about this claim, as it is adjacent to (but importantly different from) other claims that I do not support. I am clearly not claiming that the correct theory of *higher* levels is understood, which would be ludicrous. Understanding atoms and particles doesn't help much with understanding psychology or economics. I am not claiming that we understand all of particle physics; dark matter alone would be a persuasive counterexample. Nor am I claiming that we are anywhere close to the end of physics, or achieving a theory of everything. That may or may not be true, but is irrelevant to our considerations here; the correct theory of everything might require a relatively small extrapolation of our current understanding of quantum field theory, or it might ultimately involve a dramatically different and as-yet-unanticipated ontology that reduces to EQFT in some appropriate limit. Regardless, the current claim is simply that the rules

governing *one* level of reality, in a particular circumscribed regime, are fully understood. We don't know everything, and we don't know how close we are to knowing everything, but we know something, and we have a good understanding of the domain of applicability of that understanding. Finally, I am not claiming any kind of "proof" that the Core Theory suffices, even when restricted to the ELR; as is always the case in science, all we can do is offer good reasons.

This argument goes somewhat beyond a simple assertion that a particular theory does a good job at explaining certain known phenomena. The structure of quantum field theory allows us to predict the success of the model even in some circumstances where it has not yet been directly tested, given the basic assumptions on which QFT rests. It is useful to contrast the situation with that of a theory such as Newtonian gravity. The important rule there is the inverse-square law for the gravitational force,

$$\vec{F} = -\frac{GMm}{r^2}\hat{e}_r. \tag{3.1}$$

We might imagine testing this law, for example by comparing it with the motion of planets in the Solar System, and imagining that it might break down under circumstances in which it hasn't yet been tested. Indeed, by now we know that it does break down for sufficiently large values of the gravitational potential GM/r, and corrections from Einstein's theory of general relativity become important, for example in computing the precession of the perihelion of Mercury.

But there was no way of knowing ahead of time what the domain of applicability of the theory was supposed to be, other than via direct experimental test. It wasn't even possible to know what kind of phenomena would fall outside that domain. It could be (and is) when the gravitational force was strong, but it also conceivably be when the force was extremely weak (and such theories have been suggested (Milgrom, 1983)). Or when velocities were large, or when the angular momentum of the system pointed in certain directions, or when objects were made of matter rather than antimatter, or any number of other kinds of circumstances.

Quantum field theory is a somewhat different situation. Any given EFT provides its own specification of what its domain of applicability will be (as we will cover in Sect. 3.4), generally related to the energies and momenta characterizing particle interactions. As long as the basic principles are respected (quantum mechanics, relativity, locality), we can be somewhat confident that our theory is accurate within this domain, even if we haven't tested it in some specific set of circumstances. In that sense, we know a little bit more about the level of reality described by quantum field theory than we would have in other frameworks.

Our claim does have implications for how we should think about higher, emergent levels. In particular, it highlights how very radical it is to imagine that understanding complex phenomena such as life or consciousness will require departures from the tenets of the Core Theory. Such departures are conceivable, but we have good reasons to be skeptical of them. The fact that the Core Theory is so robust and difficult to modify should count strongly against placing substantial credence in that kind of strategy.

3.3 Effective Field Theory

In this section I offer a brief review of quantum field theory and the Core Theory in particular. It will necessarily be sketchy, but will serve to highlight the features that are relevant to our main point. The notion of an effective field theory will be shown to place stringent constraints on the allowed dynamics of the known fields.

Quantum field theory is a subset of, rather than a successor to, quantum mechanics. As in any quantum-mechanical theory, one has states represented by vectors in Hilbert space, an algebra of observables, and a Hamiltonian that evolves states forward in time. In practice it is more common to work with a Lagrangian L rather than a Hamiltonian; the Lagrangian is integrated over time to give an action S, which is exponentiated to provide a measure for a path integral. In a "local" QFT, the Lagrangian can be written as a spatial integral of a Lagrange density $\mathcal{L}$. The Lagrange density, Lagrangian, and action are therefore related by

$$S = \int L\, dt = \int \mathcal{L}\, d^4x, \tag{3.2}$$

where $d^4x = dt\, d^3x$ is the volume element on spacetime, and in the path-integral formalism the amplitude for a transition between two specified configurations is

$$A = \int [D\phi] e^{i S[\phi]}. \tag{3.3}$$

Here ϕ stands for all the degrees of freedom in the theory, $[D\phi]$ is a measure on the space of trajectories for those degrees of freedom, and we have suppressed an overall normalization factor.

We typically start with a classical Lagrange density—most often referred to as simply the "Lagrangian," with "density" taken as implied—and then quantize it by one of various methods. Given a set of fields, $\mathcal{L}$ is some function of those fields and their spacetime derivatives. It is often convenient to separate the terms appearing in $\mathcal{L}$ into those that are quadratic in the fields, and those that are higher-order. (Linear terms can be eliminated by re-defining fields so that such terms vanish in a stable vacuum state, while a constant term represents the vacuum energy, which we ignore in this discussion.) The quadratic terms describe the "free" theory, and higher-order terms give interactions between the fields.

The free theory can be solved exactly in Fourier space, where the field is decomposed into modes of wave vector $\vec{k}$ and wave number $k = |\vec{k}|$, corresponding to wavelength $\lambda = 2\pi/k$. These are associated with a momentum four-vector $p = (E/c, \vec{p})$, where $\vec{p} = \hbar\vec{k}$. (Henceforth we work in units where the speed of light c and the reduced Planck constant $\hbar$ are set equal to one.) For real particles, the energy satisfies $E^2 = \vec{p}^2 + m^2$, where m is the mass of the field, but for virtual particles (interior lines in Feynman diagrams), E is independent of $\vec{p}$.

In the free theory, the dynamics of any specific mode are that of a simple harmonic oscillator with frequency E. Upon quantization, the quantum state can be represented as a superposition of discrete energy levels for each mode of every field. These levels are interpreted as "particles," which is how a quantum field theory can reproduce particle physics. Fermionic fields give rise to matter particles such as leptons and quarks; bosonic fields give rise to forces, such as electromagnetism, the nuclear forces, and gravitation, as well as the Higgs field. (We are obviously skipping a great many details, including the transformation properties of the fields under symmetry transformations.)

Feynman diagrams provide a convenient graphical way of representing particle interactions. Lines entering from the left represent incoming particles, which interact by exchanging other particles, finally emerging on the right as outgoing particles. Roughly speaking, classical effects are described by tree diagrams without any internal loops, while quantum corrections are described by loop diagrams. The scattering amplitude for any specified process is obtained by adding the contributions from every possible diagram with the right incoming and outgoing particles. Figure 3.2 shows two contributions to the electromagnetic scattering of two electrons; first by the exchange of a single photon, and second by the exchange of two photons.

Each line in the Feynman diagram is labeled by the associated momentum four-vector. Momentum is conserved at each vertex, so the sum of incoming momenta must equal the sum of outgoing momenta. This condition suffices to fix the momenta of virtual particles (interior lines) in tree diagrams, but loop diagrams will have a number of undetermined momenta, one for each loop. These loop momenta are integrated over to give the contribution of that diagram to the scattering amplitude. The integration can include arbitrarily large momenta, and the resulting expressions often diverge, calling for some sort of renormalization procedure. These high-momentum (short-wavelength) divergences are known as "ultraviolet" (UV) divergences, in contrast with infrared (IR) divergences from large numbers of massless particles in the incoming or outgoing states.

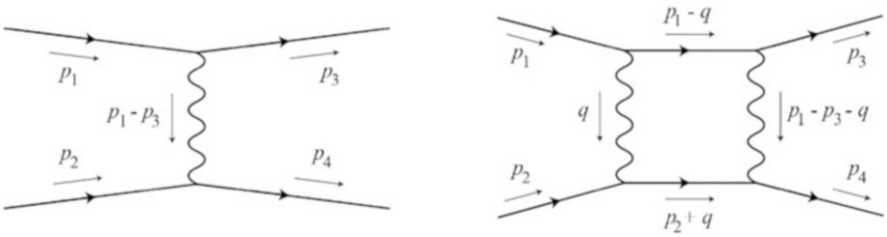

Fig. 3.2 Two Feynman diagrams for the scattering of two electrons (solid lines) by photons (waves). In the tree diagram on the left, momentum conservation at each vertex fixes the momentum of the internal photon line; in the loop diagram on the right, a free momentum q is integrated over

The modern attitude toward renormalization comes from the effective field theory program (Manohar, 2020; Rivat & Grinbaum, 2020). This approach was systematized by Wilson (Polchinski, 1984; Wilson, 1971a,b; Wilson & Kogut, 1974), though several of the important ideas had appeared earlier. Divergences come from high-energy/short-wavelength virtual particles in loops. But high energies and short wavelengths are precisely where we don't necessarily know the correct physical description. High-mass particles that are irrelevant at low energies could be important in the UV, and for that matter spacetime and the entire idea of QFT might break down at small distances.

Fortunately, as Wilson emphasized, we don't need to understand the UV to accurately describe the IR. Let us introduce by hand an energy scale Λ, the "ultraviolet cutoff." The actual value of Λ does not matter, as long as we consider incoming and outgoing momenta below that scale. In practice the effect of the cutoff is that we only integrate the momenta of virtual particles in loops up to the value of Λ, rather than all the way to infinity. This renders the loop integrals finite, though they do depend on Λ.

The physical predictions of the theory itself, however, do not depend on Λ. Rather, the original action defining the theory is replaced by an effective action $S_{\rm eff}$ for the IR modes alone. Schematically, from the path-integral perspective we have

$$A = \int [D\phi] e^{iS[\phi]} \tag{3.4}$$

$$= \int [D\phi_{\rm IR}][D\phi_{\rm UV}] e^{iS[\phi_{\rm IR},\phi_{\rm UV}]} \tag{3.5}$$

$$= \int [D\phi_{\rm IR}] e^{iS_{\rm eff}[\phi_{\rm IR},\Lambda]}, \tag{3.6}$$

where $\phi_{\rm UV}$ represents UV modes (momenta greater than Λ) and $\phi_{\rm IR}$ represents IR modes (momenta less than Λ).

Crucially, *the effective action will describe the dynamics of a local quantum field theory*, even though we have integrated out some of the degrees of freedom. Roughly speaking this is because we have eliminated modes with wavelengths less than Λ^{-1}, while considering only the dynamics of particles than can probe length scales greater than Λ^{-1}. The effective action $S_{\rm eff}$ is the integral of an effective Lagrangian $\mathcal{L}_{\rm eff}$, which can be written as a power series in the field operators. It will generally include an infinite number of terms, with arbitrarily high powers of the fields. The higher-order terms will be parameterized by coefficients that depend on the cutoff Λ, in such a way that all of the dependence on Λ completely cancels in any physical process for purely IR particles. Predictions of the effective field theory are thus independent of the arbitrary cutoff.

In presenting things this way, we have spoken as if the fundamental QFT is valid to all energies, even if we are only considering an effective theory of the IR modes. Whether or not that is the case, quantum field theory still seems to be the universal

form that physical theories take in the low-energy limit, given certain assumptions. This phenomenon of "universality" means that the most fundamental theory might feature superstrings, or discrete spacetime, or some more dramatic departure from the relativistic QFT paradigm, and still look like an EFT at low energies. Weinberg (1995) has argued that the following assumptions suffice:

- Quantum mechanics.
- Lorentz invariance.
- Cluster decomposition.
- The theory describes particle-like excitations at low energies.

(Cluster decomposition is a kind of locality requirement, that amplitudes for widely-separated scattering events be independent of each other.) This is not a rigorous result, but what Weinberg (1996) refers to as a "folk theorem." Nevertheless, it is consistent with everything we know about the universality of QFT from a variety of "ultraviolet completions," which themselves may or may not be QFTs. The explicit arguments for it only hold in the perturbative regime where fields are relatively small deviations away from the vacuum; hence, it fails to apply to strong-field phenomena like black holes.

Quantum mechanics, as is well known, is incompletely understood, or at least there is no consensus about its correct formulation. We can distinguish between the unitary-evolution part of quantum theory, where the state evolves smoothly according to the Schrödinger equation or its equivalent, and the measurement part of the theory. Unitary evolution is straightforward, but there exist multiple incompatible proposals for how we should understand the measurement process. Fortunately for our purposes, one's attitude toward the measurement problem (and fundamental quantum ontology more generally) does not affect the claim that the LPUEL are completely known. That's because all viable formulations converge, in the appropriate regime, onto the predictions of textbook quantum mechanics. Once a system in quantum superposition becomes macroscopic and entangled with its environment, it effectively "collapses" onto certain allowed measurement outcomes; details of which outcome is chosen are unpredictable aside from the Born Rule, which gives the probability of any outcome as the square of the associated amplitude within the original quantum state. That collapse may be induced by the measurement process, as in the Copenhagen interpretation; it may be truly stochastic or triggered by some physical threshold, as in objective-collapse models; it may be an artifact caused by branching of the wave function, as in Everettian quantum theory; or various other conceivable possibilities. But all of these alternatives are formulated (or at least claimed) to give rise to the same ultimate macroscopic behavior, which is all we require for our present purposes. If a new take on quantum theory predicted deviations from this textbook view, they would seemingly be accessible to experiments, which would be wonderful. Until such experimental deviations are observed, it is reasonable to stick with the textbook predictions.

None of these listed assumptions is inviolate. Quantum mechanics could be incomplete, and Lorentz invariance or locality could be merely approximate. Nevertheless, they have been tested to impressive accuracy in experiments. Without

favoring any particular stance toward the correct theory of everything describing reality might be, it makes sense to believe that the world follows the rules of effective field theory in the long-distance/low-energy perturbative regime.

These considerations are enough to eliminate one particular dependency relation that we could imagine drawing in Fig. 3.1: from everyday macro phenomena directly down to underlying reality, bypassing the QFT level. In other words, to the extent that we have good reasons to believe that the low-energy behavior of reality is accurately modeled by an effective quantum field theory, and that everyday phenomena are within that regime, we have good reason to think that there are no non-QFT phenomena characteristic of the theory of everything that are relevant for the everyday-life regime.

3.4 The Core Theory

We know more than just the general claim that low-energy physics is described by an effective quantum field theory; we know what theory it is. The Core Theory is an effective field theory that contains the well-known Standard Model of particle physics, but also quantum general relativity in the weak-field limit. The lack of a full theory of quantum gravity is a well-known outstanding issue in theoretical physics, but we have a perfectly adequate *effective* theory of quantum gravity in this regime. "Weak-field" here means essentially "small Newtonian gravitational potential GM/r," which includes everything we observe other than black holes, the very early universe, and perhaps neutron stars. It certainly covers planets in the Solar System and apples falling from trees (and for that matter gravitational waves).

In path-integral form, the theory is given by

$$A = \int_{k<\Lambda} [Dg][DA][D\psi][D\Phi]$$
$$\exp\left\{ i \int d^4x \sqrt{-g} \left[\tfrac{1}{16\pi G} R - \tfrac{1}{4} F_{\mu\nu} F^{\mu\nu} + i \bar{\psi} \gamma^\mu D_\mu \psi \right.\right.$$
$$\left.\left. + |D_\mu \Phi|^2 - V(\Phi) + \left(\bar{\psi}^i_L Y_{ij} \Phi \psi^j_R + \text{h.c.} \right) + \textstyle\sum_a \mathcal{O}^{(a)}(\Lambda) \right] \right\}. \tag{3.7}$$

This is of the general form (3.6), with an action given by a spacetime integral as in (3.2). Specific terms in the Lagrange density (large square brackets) include R for gravity, $F_{\mu\nu}F^{\mu\nu}$ for the gauge fields of the strong, weak, and electromagnetic interactions, $\bar{\psi}\gamma D\psi$ for the kinetic energy of the fermion fields, $|D\Phi|^2$ for the kinetic energy of the Higgs, $V(\Phi)$ for the Higgs potential, and $\bar{\psi} Y \Phi \psi$ for the Higgs-fermion interaction. (Interactions between gauge fields and fermions are hidden in the gauge-covariant derivative D_μ, and interactions between gravity and other fields are both there and in the overall volume element $\sqrt{-g}$ outside the brackets.) Details can be found in standard QFT texts (Peskin & Schroeder, 2015). A crucial role

here is played by the notation $k < \Lambda$ in the overall path integral, a reminder that this is an effective theory only applicable for momenta below the cutoff. The term $\sum \mathcal{O}^{(a)}(\Lambda)$ represents an infinite series of higher-order terms, each of which depend on (and in general will be suppressed by powers of) the cutoff. These terms ensure that physical predictions are independent of the cutoff value.

This is the theory that seems to underlie the phenomena of our everyday experience. The Higgs field gets a nonzero expectation value in the vacuum, breaking symmetries and giving masses to fermions. Quarks and gluons are confined into bound states such as nucleons and mesons. At low temperatures, most heavy particles decay away, leaving only protons, neutrons, electrons, photons, neutrinos, and gravitons, the latter two of which interact so weakly as to be essentially irrelevant for everyday phenomena. (Classical gravitational fields can be thought of arising from virtual gravitons. Such classical fields are relevant, but individual propagating gravitons are not.) Protons and neutrons combine into nuclei, which capture electrons electromagnetically to form atoms. A residual electromagnetic force between atoms creates molecules, and underlies all of chemistry. Finally, all of the resulting objects attract each other via gravity. Aside from nuclear reactions, everyday objects are made of electrons and roughly 254 species of stable nuclear isotopes, interacting through electromagnetic and gravitational forces.

What value for Λ should we choose? Low-energy predictions are independent of the specific value of Λ, as long as we choose it to be higher than the characteristic momentum scales of whatever processes we would like to consider. But it should also be lower than any scale at which potentially unknown physics could kick in (massive particles, restored symmetries, discrete spacetime, etc.). In practice, this means we should take Λ to be no higher than scales we have probed experimentally. For the Core Theory, we should be able to safely put the cutoff at least as high as

$$\Lambda_{\text{CT}} = 10^{11} \text{ electron volts (eV)}, \tag{3.8}$$

a scale that has been thoroughly investigated at particle accelerators such as the Large Hadron Collider. (Proton-proton collisions at the LHC have a center-of-mass energy of 10^{13} eV, but that is distributed among a large number of particles; 10^{11} eV is a reasonable value for the energy up to which individual particle collisions have been explored.) Much above that scale, and new physics is possible, and indeed many physicists are still hopeful to find evidence for supersymmetry, large extra dimensions, or other interesting phenomena.

Let us compare this to the everyday-life regime (ELR), which we are finally in position to define more precisely. The domain of applicability of an EFT is characterized by energy—more precisely, by the relative momenta of interacting particles as measured in their overall rest frame. If these momenta are all below the cutoff scale Λ, the model should be accurate. (Note that the relevant quantity is the energy per particle, not the total energy of an object, which for macroscopic objects can be quite large.) In the everyday macroscopic world, typical energies of interest are those of chemical reactions, typically amounting to a few electron volts (eV). The binding energy of an electron in a hydrogen atom is 13.6 eV, while the bond

between two carbon atoms is 3.6 eV. Bulk macroscopic motions are typically well below this energy scale; the kinetic energy of a proton in a speeding bullet is about 0.01 eV.

We might want to include nuclear reactions, such as occur in the interior of the Sun. The relevant energies are 10^8 eV or below; for example, the fusion reaction converting deuterium and tritium into helium plus a neutron releases 1.8×10^7 eV of energy. An expansive definition of the ELR, building in a bit of a safety buffer, might therefore include interactions at or below an energy of

$$E_{\rm ELR} = 10^9\,{\rm eV}. \tag{3.9}$$

All of the interactions of the particles and forces around us, and all of the radiation we absorb and admit, occurs at energies per particle lower than this value (unless we are hanging out at a high-energy particle accelerator).

The fact that $E_{\rm ELR} < \Lambda_{\rm CT}$ implies that the domain of applicability of the Core Theory encompasses the everyday-life regime. This seems to imply that not only can we list the quantum fields out of which everyday phenomena are made, but we know what their dynamics are. One loophole comes from the existence of the infinite series of higher-order terms $\sum \mathcal{O}(\Lambda)$ that inevitably appear in an effective Lagrangian. Should we be confident that they don't affect the dynamics in important ways, even at low energies?

We can gain insight by simple dimensional analysis. With $\hbar = c = 1$, energy and mass have the same units, which are the same as the units of inverse length and inverse time, and the Lagrange density has units of energy to the fourth power. Consider a real scalar field ϕ with units of energy. The part of its effective Lagrangian that contains only that field (no other fields or spacetime derivatives) is the potential energy, which takes the form

$$V_{\rm eff}(\phi) = \frac{1}{2}m^2\phi^2 + c_3\Lambda\phi^3 + c_4\phi^4 + \frac{c_5}{\Lambda}\phi^5 + \frac{c_6}{\Lambda^2}\phi^6 + \cdots . \tag{3.10}$$

Here m is the (renormalized) mass of the field, the c_is are dimensionless coefficients, and appropriate powers of the cutoff Λ appear to ensure that each term has units of (energy)^4.

The specific values of the c_is will depend on Λ (the phenomenon known as renormalization group flow), in such a way as to render physical predictions independent of Λ. But we have a "natural" expectation that these dimensionless parameters should be of order unity, rather than extremely large or small. It would be interesting to interrogate this notion of naturalness in a philosophically rigorous way, but for now we will merely note that this is indeed what happens in explicit models of EFTs where the complete UV completion is known and the parameters can be calculated as a function of Λ.

The terms in $\mathcal{L}_{\rm eff}$ can be characterized as "relevant" if they appear with positive powers of Λ (or other quantities with dimensions of energy, like m), "marginal" if they are of order Λ^0, and "irrelevant" if they appear with negative powers of Λ.

This reflects the fact that for energies well below Λ, terms with negative powers of Λ become increasingly irrelevant for making predictions. (It is these terms that are classified as "non-renormalizable.") But we've already said that our EFT is meant to be applicable only for momenta well below Λ. Therefore, our strong expectation is that these higher-order terms are indeed irrelevant for the dynamics of Core Theory fields in the ELR. (For explicit experimental constraints see Burgess et al. (1994).) The action we wrote for the Core Theory already includes all of the relevant and marginal terms that are consistent with the symmetries. We not only know what the basic fields are, but we have good reason to think that we know how they behave to very high accuracy.

3.5 New Particles and Forces

If we believe we understand the dynamics of the known fields of the Core Theory, the other way that model could fail to completely account for everyday phenomena—without leaving the EFT paradigm entirely—is if there are unknown fields that could play a subtle but important role. We can distinguish between three ways this could happen.

- A new field could show up as virtual particles mediating a new kind of interaction between the known fields. However, this would essentially modify the low-energy effective action (3.7) of the Core Theory. This would have no observable effects unless the results deviated significantly from our effective-field-theory expectations, and as we have noted there are good constraints on any such possibility. So we will not consider this alternative in detail.
- A field could give rise to new long-lived particles that played a distinct dynamical role in macroscopic phenomena, much like electrons, protons, and neutrons do. Such a particle could be ambient in the universe, much like dark matter but possibly with a lower overall energy density. Perhaps a particle of this form participates in the neurochemical processes of conscious creatures (Pullman, 2000).
- A weakly-interacting bosonic field could condense to give a classical force field, what physicists think of as a "fifth force." Such a force could conceivably induce interactions between neurons, or even between different brains, as two vivid examples.

Let's consider these last two possibilities in turn.

In contemplating the existence of novel ambient particles, it is useful to compare with the case of neutrinos, which are known to exist. There are a lot of neutrinos in the universe; the flux near Earth, from both the cosmic neutrino background and solar-generated neutrinos, is of order 10 trillion neutrinos per square centimeter per second. But they interact with ordinary matter quite weakly (literally through the "weak interactions" of the Standard Model), so much so that of the order 10^{21} neutrinos that pass through a typical human body in a typical lifetime,

approximately one of them will actually interact with the atoms in that body. Any hypothetical new particle would have to have substantially higher interaction strength with ordinary matter in order to play a role in everyday phenomena.

One way of constraining such new particles is by simply trying to create them at particle accelerators. The QFT property of crossing symmetry guarantees that such searches are feasible. Consider a new particle X that interacts with electrons through some new force, mediated by a new field Y; within the EFT paradigm, something along these lines would be necessary for X to affect everyday objects. In Feynman-diagram language we can represent that as an incoming electron and X, which interact via virtual Y exchange and then continue on. Crossing symmetry implies that the amplitude for such an interaction will be related to that obtained by rotating the diagram by ninety degrees, and interpreting particles going backward in time as antiparticles. Hence, this scattering amplitude is related to the amplitude for an electron and positron (anti-electron) to annihilate into a Y, which then decays to an X and an anti-X, as shown in Fig. 3.3.

Fortunately, colliding particles together and studying what comes out is particle physicists' stock in trade. Our X particle must be electrically neutral and invisible to the strong nuclear force, otherwise it would interact very noticeably and have been detected long ago. It therefore won't leave a visible track in a particle detector, but there are indirect methods for constraining its existence. For example, new particles give other particles new ways to decay, decreasing their lifetime and therefore increasing the width of energy distribution of particles into which they decay. (This can be thought of as a consequence of the energy-time uncertainty principle; faster decay implies more uncertainty in energy.) The decay width of the Z boson was measured to high precision by the Large Electron-Positron Collider, a predecessor to the Large Hadron Collider at CERN. Results are usually quoted in terms of the number of "effective neutrino species," although the principle applies to non-neutrino particles as well. (Even if X coupled to quarks and not to electrons, it would still be produced by interactions with virtual quarks.) There are three conventional neutrino species in the Core Theory, and the LEP measurement came in at 2.9840 ± 0.0082 (Mele, 2015). We can interpret this as saying that there are no unknown particles with masses less than half that of the Z (about 4×10^{10} eV)

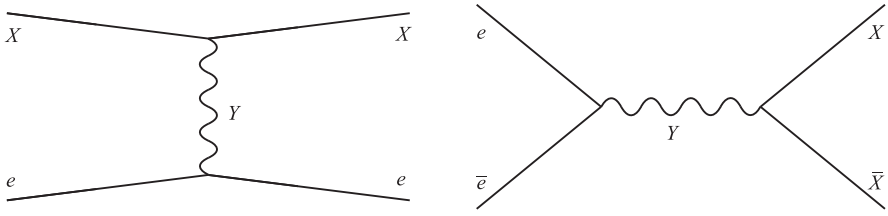

Fig. 3.3 Crossing symmetry relates the amplitudes for these two processes, an interaction of a new particle X with an electron e via a mediator Y, and annihilation of an electron/positron pair into an X/anti-X. Any new particle that interacts with ordinary matter can therefore be created in particle collisions

that interact with Core Theory fermions with an interaction strength greater than or equal to that of neutrinos.[1]

Heavier X particles can also be constrained, and other measurements also provide limits (Aad et al., 2020; Acciarri, 1999; Fox et al., 2012). If X particles are extremely heavy, say over 10^{11} eV, they would be out of reach of current particle accelerators. But if such particles are ambient, there is a limit on how abundant they can be, given by the dark-matter density. (If new stable particles have more mass density than dark matter, they would be ruled out by astrophysical measurements.) So as not to have more mass density than dark matter, an ambient particle of mass m must have a number density lower than about $(3 \times 10^{11}$ eV/m) per liter in the Solar System. It is hard to imagine such dilute particles being relevant for everyday dynamics.

We have noted that neutrinos barely interact with ordinary matter at all; any hypothetical new ambient particle that would be relevant to the behavior of macroscopic objects would have to interact much more strongly than that. Particle-physics constraints imply that there are no such particles. New particles may certainly exist, but they must be either short-lived, weakly-interacting, or extremely rare in the universe. We can therefore conclude that unknown ambient particles do not play a role in accounting for phenomena in the everyday-life regime.

The other reasonable option is the existence of a bosonic field that couples weakly to individual particles, so that direct searches for the boson would be fruitless, but that is sufficiently low-mass that it can accumulate to give rise to a macroscopic force field. (The range of a field is inversely proportional to its mass, with $r[\text{cm}] \sim 2 \times 10^{-5}/(\text{m}\,[\text{eV}])$.) For our purposes here we could define "macroscopic" as larger than one micrometer; the average cell in a human body is between 10 and 100 micrometers in diameter.

Gravity itself is an example of a field whose quanta are undetectable but that gives rise to a macroscopic force. Individual gravitons couple far too weakly to be detected, but the net gravitational force sourced by matter in the Earth is enough to keep us anchored to the ground, because the gravitational field is infinite-range (gravitons are massless) and every particle contributes positively to the force. Gravity is nevertheless extremely weak; the gravitational force between two typical human bodies separated by a distance d is less than 10^{-7} the electromagnetic force between two individual protons at the same separation. To be generated by human-sized (or smaller) objects, and yet have a noticeable impact on the dynamics of the macroscopic world, a new force would have to be enormously stronger than gravity. This seems unlikely at first glance, as we would presumably have noticed such a force. But it's conceivable that it couples only to certain combinations of particles (rather to everything, as gravity does), and that it has a macroscopic but finite range,

[1] One subtlety is that the electron-X interaction could be enhanced if the two particles exchanged a large number of virtual Ys; something similar happens in ordinary electromagnetism. But that would require the Y itself to be a very light particle, and then it would contribute the number of effective neutrino species bounded by LEP.

so that it doesn't affect celestial dynamics or apples falling from trees. It's therefore worth examining the possibility more carefully.

Fortunately, there aren't that many different ways in which a fifth force can couple to ordinary matter. Within the framework of low-energy effective field theory, we can think of the source of the new force as some linear combination of electrons, protons, and neutrons. The available parameter space can be constrained by measuring the forces between macroscopic objects of substantially different chemical compositions. We don't need to be too precise about the results here, as a rough guide is more than adequate for our purposes. From a variety of experimental and astrophysical techniques, stringent bounds have been placed on the possible existence of new long-range forces (Adelberger et al., 2009); the results are summarized in Fig. 3.4.

It is clear from examination of this plot that for ranges greater than 10^{-4} m (100 micrometers), any new force must be weaker than gravity, and at 10^{-3} m and above the limits are better than 10^{-3} gravity. Given how weak gravity itself is between human-sized objects, this definitively rules out the possibility that such forces are important for dynamics in the ELR. At shorter ranges the limits deteriorate, both because the magnitude of the force between small test objects is smaller and harder to measure, and (more importantly) because it becomes harder to eliminate possible contamination from residual electromagnetic forces. For precisely this reason, such forces will also be irrelevant for macroscopic dynamics. At one micrometer, a force

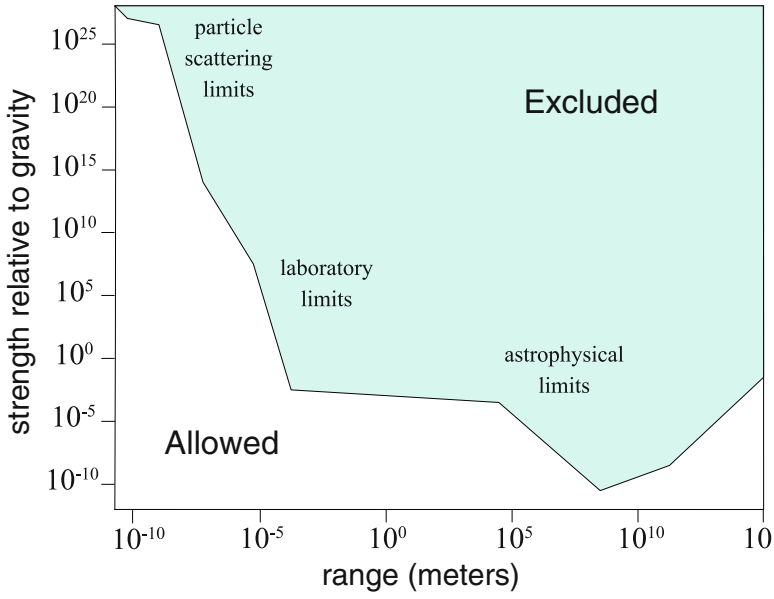

Fig. 3.4 Limits on a new fifth force, in terms of its strength relative to gravity, as a function of its range. Adapted from data collected in Adelberger et al. (2009). This is a rough reconstruction; see original source for details

10^9 times gravity would be allowed, but that is only 10^{-27} times the strength of electromagnetism. Even with substantial cancellations between positive and negative charges, residual electromagnetic forces will overwhelm a fifth force at these ranges. All the way down at atomic scales, $\sim 10^{-10}$ m, any new force must still be less than 10^{-6} the strength of electromagnetism.

We therefore conclude that, within the framework of effective field theory, there is no room for unknown fields or unanticipated dynamics to play a role in accounting for macroscopic phenomena in the everyday-life regime. There can be, and very likely are, more fields yet to be discovered, but they must either be extremely dilute in the universe so that we essentially never interact with them, or so weakly coupled to ordinary matter that they exert essentially no influence. Quantum field theory might not, and probably is not, the correct framework in which to formulate an ultimate theory of everything, but given certain plausible assumptions low-energy physics will nevertheless be accurately modeled by an EFT, so everyday phenomena do not depend directly on deeper levels, only through the Core Theory. There is much of physics that we don't know, and it is entirely unclear how close we are to achieving a fundamental theory of nature. But we do understand the laws of physics underlying everyday phenomena as described at one particular level of reality, that of effective quantum field theory.

3.6 Discussion

I have argued that we have good reasons to believe that everyday-life phenomena supervene on the Core Theory, and not on as-yet-undiscovered particles and forces or on new principles at more fundamental levels. The argument relies on an assumption that the world is entirely physical, and that there is a level of reality accurately described by an effective quantum field theory. Then the general properties of quantum field theory, plus known experimental constraints, lead us to the conclusion that the Core Theory suffices.

If this package of claims—physicalism, EFT, Core Theory—is correct, it has a number of immediate implications. There is no life after death, as the information in a person's mind is encoded in the physical configuration of atoms in their body, and there is no physical mechanism for that information to be carried away after death. The location of planets and stars on the day of your birth has no effect on who you become later in life, as there are no relevant forces that can extend over astrophysical distances. And the problems of consciousness, whether "easy" or "hard," must ultimately be answered in terms of processes that are compatible with this underlying theory.

Less obviously, our understanding of the Core Theory has implications for the development of technology. Historically, progress in fundamental physics (as it was defined at the time) has often had important technological implications, from mechanics and electromagnetism to quantum theory and nuclear physics.

That relationship has largely evaporated. The last advance in fundamental physics (defined in a modern context as new particles or forces or dynamics at the quantum-field level) to be put to use in technology was arguably the discovery of the pion in 1947. Since then, technological development has depended on increasingly sophisticated ways of manipulating the known particles and forces in the Core Theory. This is likely to be the case for the foreseeable future; the kinds of new particles remaining to be discovered either require multi-billion-dollar particle accelerators to produce (and even then they decay away in zeptoseconds), or they interact with ordinary matter so weakly as to be essentially impossible to manipulate in useful ways. It is hard to imagine technological applications of such discoveries. Even quantum computing, which has involved important conceptual breakthroughs, makes use of the same underlying physical matter and laws that have been known for well over half a century.

Needless to say, the claim that we fully understand the laws of physics underlying everyday life might very well be incorrect, even if there are good reasons to accept it. It is easy enough to list some potential loopholes to the argument, ways in which the claim might fail to be true by going outside the EFT framework.

- Violations of locality. In the context of an EFT, locality of the Hamiltonian implies that the electromagnetic or gravitational fields (or unknown fifth-force fields) produced by an object are simply the net fields produced by each of the constituent particles individually. Outside the EFT paradigm, we could imagine forces that depend non-locally on sources, so that whether or not a force is produced would depend on the specific arrangement of particles within it. (This is completely distinct from the non-locality associated with quantum measurements.) Such a force might not be produced by a collection of electrons, protons, and neutrons in the form of a cantaloupe, for example, but be produced by the same particles when they are in the form of a human brain. To the best of my knowledge, this possibility has not been investigated carefully (and to be honest, there is not a lot of motivation for it).
- Quantum wave function collapse. In conventional quantum mechanics, the probability of a measurement outcome is given by the absolute-value squared of the corresponding amplitude of the wave function (the Born Rule). Other than that, the process is thought to be entirely random, with no structure other than that statistical rule. But perhaps it is not, and quantum systems evolve in subtle and specific ways to bring about particular outcomes. This scenario has been studied, typically in the context of trying to attain a better understanding of consciousness (Penrose, 1989; Chalmers & McQueen, 2014).
- Departures from physicalism. Everything we have said presumes from the start that the world is ultimately physical, consisting of some kind of physical stuff obeying physical laws. There is a long tradition of presuming otherwise, and if so, all bets are off. The well-known issue is then how non-physical substances or properties could interact with the physical stuff.

This list is not meant to be exhaustive, but provides a flavor of the options available to us.

The reasons for denying the claim advanced in this paper, and going for one of the above loopholes instead, generally arise from a concern that the physical dynamics of the Core Theory cannot suffice to account for higher-level phenomena, whether the phenomenon in question is life after death or the experience of qualia. Our considerations do not amount to an airtight proof (which would be essentially impossible), but they do highlight the challenge faced by those who think something beyond the Core Theory is required. The dynamics summarized in Eq. (3.7) are well-defined, quantitative, and unyielding, not to mention experimentally tested to exquisite precision in a wide variety of contexts. Given a quantum state of the relevant fields, it accurately predicts how that state will evolve. Skeptics of the claim defended here have the burden of specifying precisely how that equation is to be modified. This would necessarily raise a host of tricky issues, such as possible violations of conservation of energy or non-unitary evolution of the quantum state (even when unmeasured and unentangled). A simpler—though still extremely challenging—alternative is to work to understand how those dynamics give rise to the emergent levels of reality in our macroscopic world.

Acknowledgments It is a pleasure to thank Jenann Ismael, Ira Rothstein, Charles Sebens, and Mark Wise, as well as an anonymous referee, for helpful comments on a draft version of this manuscript. This research is funded in part by the Walter Burke Institute for Theoretical Physics at Caltech, by the U.S. Department of Energy, Office of Science, Office of High Energy Physics, under Award Number DE-SC0011632, and by the Foundational Questions Institute.

References

Aad, G., et al. (2020). Dijet resonance search with weak supervision using $\sqrt{s} = 13$ TeV pp collisions in the ATLAS detector. *Physical Review Letters, 125*(13), 131801.

Acciarri, M. E. A. (1999). Single and multi-photon events with missing energy in e-$\bar{e}$ collisions at $\sqrt{s}$ = 189 Gev. *Physics Letters B, 470*(1–4), 268–280. https://doi.org/10.1016/S0370-2693(99)01286-1

Adelberger, E., Gundlach, J., Heckel, B., Hoedl, S., & Schlamminger, S. (2009). Torsion balance experiments: A low-energy frontier of particle physics. *Progress in Particle and Nuclear Physics, 62*(1), 102–134. http://www.sciencedirect.com/science/article/pii/S0146641008000720

Burgess, C., Godfrey, S., König, H., London, D., & Maksymyk, I. (1994). Model-independent global constraints on new physics. *Physical Review D, 49*(11), 6115.

Carroll, S. M. (2017). *The big picture: On the origins of life, meaning, and the universe itself*. Westminster: Penguin.

Chalmers, D., & McQueen, K. (2014). Consciousness and the collapse of the wave function. In *Quantum Mechanics and Consciousness*.

Fox, P. J., Harnik, R., Kopp, J., & Tsai, Y. (2012). Missing energy signatures of dark matter at the LHC. *Physical Review D, 85*(5). https://doi.org/10.1103/PhysRevD.85.056011

Manohar, A. V. (2020). *Introduction to effective field theories*. *Les Houches Lecture Notes* (Vol. 108). Oxford: Oxford University

Mele, S. (2015). The measurement of the number of light neutrino species at LEP. In *60 Years of CERN Experiments and Discoveries* (pp. 89–106). London: World Scientific.

Milgrom, M. (1983). A modification of the newtonian dynamics as a possible alternative to the hidden mass hypothesis. *The Astrophysical Journal, 270*, 365–370.

Penrose, R. (1989). *The emperor's new mind: Concerning computers, minds, and the laws of physics*. Oxford: Oxford University Press.

Peskin, M. E., & Schroeder, D. V. (2015). *An introduction to quantum field theory*. Boca Raton: CRC Press.

Polchinski, J. (1984). Renormalization and effective lagrangians. *Nuclear Physics B, 231*, 269–295.

Pullman, P. (2000). *The Amber spyglass*. Oxford: Scholastic/David Fickling Books.

Rivat, S., & Grinbaum, A. (2020). Philosophical foundations of effective field theories. *The European Physical Journal A, 56*(3), 1–10.

Weinberg, S. (1995). *The Quantum Theory of Fields* (Vol. 1). Cambridge: Cambridge University Press.

Weinberg, S. (1996). What is quantum field theory, and what did we think it is? In *Conference on Historical Examination and Philosophical Reflections on the Foundations of Quantum Field Theory* (pp. 241–251).

Wilczek, F. (2015). *A beautiful question: Finding nature's deep design*. Westminster: Penguin.

Wilson, K. G. (1971a). Renormalization group and critical phenomena. 1. Renormalization group and the Kadanoff scaling picture. *Physical Review B, 4*, 3174–3183.

Wilson, K. G. (1971b). Renormalization group and critical phenomena. 2. Phase space cell analysis of critical behavior. *Physical Review B, 4*, 3184–3205.

Wilson, K., & Kogut, J. B. (1974). The Renormalization group and the epsilon expansion. *Physics Reports, 12*, 75–199.

Chapter 4
Against Levels of Reality: The Method of Metaphysics and the Argument for Dualism

Michael Esfeld

Abstract This paper has three objectives: (i) arguing against levels of reality by employing the Lewis-Jackson method for doing metaphysics, also known as the Canberra plan; (ii) showing how this method renders the idea of levels of reality incoherent, but nevertheless leaves the conceptual space open for dualism; (iii) sketching out a concrete proposal for a dualism of mind and matter that relies on normativity and that employs ontic structural realism.

4.1 Against Levels of Reality: The Canberra Plan

Consider how Jackson (1994, p. 25) describes the task of metaphysics:

> Metaphysics, we said, is about what there is and what it is like. But of course it is concerned not with any old shopping list of what there is and what it is like. Metaphysicians seek a comprehensive account of some subject matter – the mind, the semantic, or, most ambitiously, everything – in terms of a limited number of more or less basic notions. In doing this they are following the good example of physicists. The methodology is not that of letting a thousand flowers bloom but rather that of making do with as meagre a diet as possible. ... But if metaphysics seeks comprehension in terms of limited ingredients, it is continually going to be faced with the problem of location. Because the ingredients *are* limited, some putative features of the world are not going to appear explicitly in the story. The question then will be whether they, nevertheless, figure implicitly in the story. Serious metaphysics is simultaneously discriminatory and putatively complete, and the combination of these two facts means that there is bound to be a whole range of putative features of our world up for either elimination or location.

This is a paradigmatic statement of what is known as the Canberra plan: metaphysics is ontology, answering the question of what there is in such a way that something is admitted as primitive and that it is then shown how everything else that exists is included in what is endorsed as primitive (location). This implies that the

M. Esfeld (✉)
Department of Philosophy, University of Lausanne, Lausanne, Switzerland
e-mail: Michael.Esfeld@unil.ch

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_4

propositions that describe the world in terms of the primitive notions entail all the other true propositions about the world. However, what is primitive does not constitute a fundamental level. There are no levels. There is only one layer of reality that is described in terms of the primitive notions. It is then shown how everything that exists is located in this layer and how its description is entailed by the description of this lawyer by the primitive notions. If the primitive notions are only physical ones, the result is a position that Shenker (2017) has aptly characterized as "flat physicalism". Hence, this methodology for metaphysics hinges upon the ability to define a precise set of primitive notions: there is no endless way of going down to ever further notions that are primitive relative to other notions.

Let us consider a concrete example of how this can go. What is an ontology that is minimally sufficient to account for our scientific as well as our common sense knowledge about the natural world in the spirit of scientific realism? In Esfeld and Deckert (2017, ch. 1), it is argued that an ontology of the natural world defined in terms of the following two axioms is an answer to this question (one answer, not the only possible answer):

1. *There are distance relations that individuate simple objects, namely matter points (point particles).*
2. *The matter points are permanent, with the distances between them changing.*

The reason for singling out the distance relation is that it is the first and foremost candidate for the world-making relation, at least insofar as the natural world is concerned: all and only those objects that stand in a distance to each other make up a world. In other words: distance provides for extension, and extension is generally admitted as being characteristic of the natural world (*res extensa* in Descartes' terms).

If distances are indispensable anyway, one can employ them to individuate the basic objects of which everything else in the physical world is composed: what distinguishes physical objects from one another are their relative positions in the configuration of matter of the universe. No commitment to intrinsic essences is called for; these would not be able to distinguish individual objects anyway. Since the world is not static, but change happens, a second axiom is mandatory that captures change, which then consists in change in the relative distances among the point objects. No commitment to absolute space and time is required: space is the order of what coexists, namely a configuration of matter points individuated by their relative distances; time is the order or measure of change, as Leibniz (1890) maintains, notably in his third and fourth letter to Newton and Clarke. Consequently, the specific notions endorsed as primitive are the ones of matter points, distances and change of the distances. Esfeld and Deckert (2017) provide a detailed account of how one can reconstruct physics – from classical via relativistic to quantum physics, including quantum field theory – on this parsimonious basis, "making do with as meagre a diet as possible" as Jackson puts it. Esfeld (2020, ch. 1) further elaborates on the metaphysical aspects of this view.

Nonetheless, this is just one example of how the Canberra plan can be put to work. Even if it can be argued that these two axioms are sufficient to capture the existing natural science, future progress in science may require going over the metaphysical books. In short, any attempt to implement the Canberra plan depends on the actual science at the time the attempt is made.

Why should one endorse this stance given the scientific knowledge that we have at our disposal? In a nutshell, the argument is this one: consider two possible worlds that agree on the spatio-temporal arrangement of matter, that is, agree on the relative positions of the material objects all the time, that is, throughout all their change. Any such worlds are indiscernible by any scientific means. By the same token, if a theory gets the spatio-temporal arrangement of matter right (that is, the arrangement of fermionic matter according to contemporary physics, as e.g. Bell (1987, p. 175) points out), it has got everything right that can ever be checked in scientific experiments (see also Maudlin, 2019, pp. 49–50). Two theories that agree on the spatio-temporal arrangement of the basic discrete objects cannot be distinguished by any empirical means, whatever else they may otherwise say and disagree on. Agreement in the spatio-temporal arrangement of matter means agreement in the relative distances of objects that are characterized by these distances only. Whatever else a theory may attribute to these objects (such as masses, charges and the like) and whatever else it may pose (such as fields, waves as well as wave functions and the like) is accessible to scientific investigation only in terms of changes in the relative distances among discrete objects with these changes then being conceptualized in terms of attributing masses, charges, wave functions, etc. to these objects. However, reifying these magnitudes to something that the objects possess in and of themselves over and above standing in relations of relative distances that individuate them runs into the type of objection that Leibniz addresses to Newton against absolute space and time, namely the commitment to a surplus structure in the ontology that leads to differences in possible worlds that make no empirical difference and hence no difference that can be investigated by any scientific means.

Hall (2009, § 5.2) makes this point in the following way:

> ... the primary aim of physics – its first order business, as it were – is to account for *motions*, or more generally for change of spatial configurations of things over time. Put another way, there is one Fundamental Why-Question for physics: Why are things located where they are, when they are? In trying to answer this question, physics can of course introduce *new* physical magnitudes ...

This suggests that the new physical magnitudes – that is, all the variables beyond the primitive variable of relative positions – can be introduced in terms of the role that they play for the change in the relative positions of the discrete physical objects. In other words, all there is to them is their function in the account of the evolution of the relative positions of objects that a theory formulates. This means that these variables are located in the motion of the objects instead of being something over and above relative positions and their change. To put it differently, propositions that employ terms such as “mass”, “charge”, etc. are true. However, their truthmaker is not an intrinsic mass or charge, etc. that objects have over and above relative

positions; their truthmaker is the way in which the objects move, that is, the overall change in their relative positions. To put it in a nutshell, some objects are electrons – that is, have negative charge – because they move electronwise, that is, behave like electrons.

This stance has become known as *Super-Humeanism* (see Esfeld & Deckert, 2017, ch. 2.3; Esfeld, 2020, ch. 2). It goes beyond the Humean metaphysics set out, for instance, in Lewis (1986, introduction) in that it defines the Humean mosaic only in terms of distance relations that individuate simple objects. Hence, instead of the natural, intrinsic properties that Lewis poses, there is only one natural relation that is the world-making relation and that individuates the objects. The stock objections against Lewis's Humean metaphysics from quidditism and humility are thus avoided, because there are no natural, categorical properties.

Indeed, functionalism is the solution to the problem of location (or placement, to use the term of Price, 2004). The ontology is in any case given by the notions that are admitted as primitive – in the case at hand, the notions of "matter points", "distances" and "change of distances". One then defines everything else in terms of its function in the sense of the role that it plays for that change. That functional role is realized by the ontology as defined by the primitive notions. Consequently, everything else is thereby located in that ontology and its description is entailed by that ontology, given the functional definitions.

Let us review some stock examples to illustrate this method. Consider water. As we know from scientific investigation, there is no primitive water stuff in the world. Science superseded the ancient view of the four elements earth, water, air and fire. But, of course, there is water in the world: there are things in the world that fulfil the functional role of appearing odourless, colourless, being thirst-quenching through the change in the motion of the parts of our bodies that they bring about. These are configurations of H_2O molecules. Thus, by defining water in terms of its thirst-quenching role – that is, its role for certain motions in our bodies –, we locate water in the ontology of particles that move: certain particle configurations, moving in certain characteristic ways, *are* water.

By the same token, there is no *élan vital*, no *sui generis* life stuff or causal power; but there are organisms in the world. The functional role that defines what it is to be alive in terms of certain characteristic motions such as reproduction and adaptation to the environment is realized by certain configurations of molecules, as we know since the advent of molecular biology in the twentieth century. Again, this means that certain particle configurations, moving in certain particular ways, *are* organisms. Life thus is located in certain particle configurations.

Furthermore, according to physicalism, there are no *sui generis* minds; but there are mental states defined by certain functional roles, which in the end are functional roles for the behaviour and thus the bodily motions of persons, realized by certain neuronal configurations. This functionalist stance goes back to Lewis (1966) and has been forcefully argued for by Kim (1998) and others. Again, this means that certain particle configurations – in this case, certain neuronal configurations –, moving in certain particular ways, *are* minds.

The point of Super-Humeanism is to apply this method of location via functional definitions not only to the objects of the special sciences, but already within physics. Consider gravitation: the motion of the objects in the world manifests some salient patterns or regularities. Arguably the most striking of these patterns is mutual attraction. This pattern applies everywhere and at every scale in the universe, from atoms to apples falling from trees and to planetary motion, such as the motion of the Earth around the Sun. This stable pattern enables us to introduce the notion of gravitational mass in order to represent this regular motion: gravitational mass is defined in terms of its function for particle motion, namely the role of mutual attraction. Already Mach, for instance, brings this functional definition of mass out in his *Science of mechanics* when saying that "The true definition of mass can be deduced only from the dynamical relations of bodies" (Mach, 1919, p. 241). Russell (1912) makes the same point in his famous paper on the notion of causation.

All the evidence that we have are the dynamical relations of bodies – that is, their motions; these relations manifest certain stable patterns, such as attractive motion. To represent these patterns in a theory, physicists introduce various parameters that are defined by their function for the particle motion. These may be parameters that are attributed to the individual objects and that remain fixed, such as mass, charge, spin, etc., parameters that evolve in time such as energy or a wave function, etc. as well as constants of nature. In short, on Super-Humeanism, not only the laws, but also the dynamical parameters that a theory employs over and above the primitive parameter of relative positions as well as the geometry of space-time come in as a package in order to accomplish the best system – that is, a representation of the motion of matter that strikes the best balance between being simple and being informative.

Lewis (1986, introduction) employs the notion of supervenience: Humean supervenience is the claim that everything else supervenes on the Humean mosaic of matter in motion as defined by the primitive notions. On Super-Humeanism, this is the configuration of point particles of the universe that are individuated by their relative distances and the change in distances. However, for Lewis and Jackson, supervenience means identity as well as *a priori* entailment of the propositions describing everything else by the propositions that describe the world in terms of the primitive notions, given functional definitions of everything else. This is what the analytic, reductive functionalism that is set out in Lewis (1966, 1970, 1972) amounts to. It is therefore recommendable to stick to the notion of identity, because it is simple and clear, and to the method of location through functional definitions, because it is precise.

Identity is symmetrical, whereas supervenience is not. If, for instance, certain particle configurations are identical with the water that there is in the universe by playing the water role against normal background conditions, then the water that there is in the universe is identical with certain particular particle configurations. Nonetheless, despite being symmetrical, this identity amounts to an ontological reduction, which is not symmetrical: everything is particles and their configurations (that is, reduced to particles and their configurations), whereas only some specific particle configurations are water, organisms, etc. Hence, the notion of identity is

clear and simple and, yet, does the service for which it is employed here: it expresses how everything else is located in what is described by the primitive notions of there being point particles individuated by relative distances and the change in these distances.

In the current literature, the notion of supervenience is often replaced with the one of grounding (see e.g. the essays in Correia & Schnieder, 2014). Applied to our context, grounding is to say that the configuration of matter as defined by relative distances individuating point particles and their change grounds everything else in the sense that it is a sufficient condition for everything else. However, grounding is not identity. The concept of grounding expresses a correlation between something that is designated as fundamental and all the rest and accords ontological priority to what is designated as fundamental. But this correlation, however robust it may be, remains a brute fact. Grounding does not explain anything. By contrast, the method of location via functional definitions yields an explanation: providing a functional definition of something and on that basis showing how that something is realized by what is admitted as primitive as described by the primitive notions answers the question why there is that something and how it comes in given what is admitted as primitive. To come back to one of the stock examples, saying that water is grounded in H_2O molecules does not answer the question why there is water. Providing a functional definition and on that basis showing how H_2O molecules realize the water role in the world so that they are identical with water explains why there is water.

This, then, is the argument against levels of reality: locating everything else in an ontology defined by a minimal set of primitive notions explains everything else by showing how it is identical with something in that ontology. If one renounces on identity as embedded in this conception of location through functional definitions, one is left with brute correlations among a basic level of reality and higher levels of reality, whatever notions one may employ to designate that basic level as fundamental (supervenience, grounding, etc.).

That notwithstanding, there are obviously new features coming up in the evolution of the universe, that is, features that are limited to specific places and times, such as the formation of water molecules, or the development of organisms, etc. and that are in this sense emergent features of the universe. However, in science, these features are explained in terms of the dynamical laws that apply everywhere in the universe plus special initial conditions, which, again, are special initial conditions of the universe in the last resort. For instance, what is known as the past hypothesis, stating that the initial particle configuration of the universe is one that implements a very low entropy, is crucial in order to give a scientific explanation of why organisms evolve at certain times and places in the universe. More precisely, such a scientific explanation tells us why particle configurations evolve that realize organisms, etc., and the method of location via functional definitions tells us why these configurations are organisms, etc. Hence, they are not new ontological features of the universe: by means of such an explanation, they are located in the particle configuration and its evolution. Thus, far from being opposed to reduction, emergent features in the sense of new features coming up in the evolution of the universe just are the object to which the methodology of location through functional definitions

is designed to apply in the first place (although it applies also already to universal physical features such as mass and charge).

4.2 The Methodology of the Canberra Plan Beyond the Natural Sciences

The Canberra plan provides a clear roadmap for both ontology and epistemology. As regards ontology, the task is to set out the ontology in terms of a few notions that are admitted as primitive and then to show how the ontology thus defined includes everything, because all the things that are not described explicitly by the primitive notions are located in that ontology through functional definitions. As regards epistemology, all further notions apart from the primitive ones that define the ontology come in through a definition in terms of a functional role for the behaviour of what is described by the primitive notions. In general, given the description of the world in terms of the primitive notions and such functional definitions of everything else, the propositions describing everything else are entailed by the propositions that describe the world in terms of the primitive notions. The multiple realizability of functional roles does not infringe upon these entailment relations: the issue are sufficient physical conditions, defined in terms of the primitive notions, for these roles to be realized, never necessary and sufficient conditions and thus never biconditionals; this is made clear, for instance, in Chalmers (1996, pp. 42–51) on reductionist explanations, in Esfeld and Sachse (2011, ch. 5) on conservative reductionism and in Hemmo and Shenker (2015) on the emergence of macroscopic regularity.

Thus, on the proposal sketched out in the preceding section, everything in the physical world is identical with a configuration of matter points that is characterized only by the relative distances among the matter points and the change in these distances. Consequently, "matter points", "distances" and "change of distances" are the primitive notions employed to describe the world. The task then is to find out which configurations of matter points are water, genes, organisms, etc. given functional definitions of these things in terms of the role that they play for the motion of matter, that is, in the last resort, the evolution of the distance relations among the matter points.

The Canberra plan can obviously be applied beyond the domain of the natural sciences. The crucial issue is the functional definition of the relevant concepts in terms of their functional role for, in the last resort, particle motion. Consider mental concepts: there is no question any more today of behaviourism, that is, of defining mental concepts directly in terms of a role for the bodily motions of persons. Nonetheless, functionalism in the philosophy of mind is the successor of behaviourism, as pointed out, for instance, by Lewis (1966, section III). The functional definition of each single mental concept can include other mental concepts; but in the end, the functional definition of the whole cluster of mental

concepts is one in terms of their causal role for the behaviour of the person, that is, for the change in the relative positions of the particles making up the person's body and its environment.

This is just a matter of definition. One can simply stipulate that everything else be defined in terms of a causal role for, in the last resort, particle motion. The crucial issue is whether such a definition is convincing, that is, whether it captures the being or the essence of the targeted things. There is no such debate as far as physical dynamical parameters such as mass, charge, etc. are concerned: they are introduced in physics in terms of the role that they play for the particle motion. Generally speaking, functional definitions of this kind are undisputed in the natural sciences. It would be odd, for instance, to postulate a heat stuff to account for thermodynamical phenomena, since these can be defined functionally in terms of changes in molecular motion. By the same token, it would be odd to postulate an essence of water over and above interacting H_2O molecules, or an *élan vital* to capture organisms and their reproduction. Since the advent of molecular biology, the evolution of organisms and their reproduction can be accounted for in terms of molecular biology so that functional definitions in terms of causal roles for, in the last resort, particle motion are vindicated. There is no explanatory gap here between descriptions in terms of molecular motion and descriptions in terms of heat, water, genes, etc.

However, there is a debate when it comes to the mind. One can doubt that functional definitions seize the qualitative aspects of conscious experience (so called qualia, giving rise to what is known as the hard problem of consciousness). Furthermore, one can doubt whether functional definitions in terms of causal roles for, in the last resort, behaviour and thus particle motion capture the rational side of the mind, which includes thoughts, intentions to act and in general deliberations about what one should think and do.

The Canberra plan remains silent on the question as to what extent such functional definitions are successful. It limits itself to setting out a clear methodology for metaphysics or ontology: first, one expresses the ontology in terms of a minimal set of notions that are endorsed as primitive – such as "matter points", "distances" and "change of distances" on the proposal discussed in the preceding section. Accordingly, the ontology endorsed as primitive – that is, as not derived from anything else – then is the one of matter points individuated by the distances among them and the change of these distances. As regards everything else that does not figure explicitly in this ontology, there then are the following three possibilities:

- *Location* in the ontology through functional definitions in terms of a role that is realized in the ontology as defined by the primitive notions. This applies to everything in the domain of the natural sciences.
- *Elimination*: The thing in question does in fact not exist. For instance, it would be futile to seek to locate witches in the ontology, because there are no witches. It is an error to think that certain things (people for that matter) are witches.
- *Further primitives*: If something can neither be located in the ontology as defined by the primitive notions, because a functional definition of it in terms of a role for the behaviour of that ontology does not seize its being or essence, nor

be eliminated, because there is overwhelming evidence of its existence, then that something has to be admitted as a further primitive. Hence, the ontology originally posed as primitive has to be enlarged.

The reasoning is this one: for everything that is a candidate for something real, the thing in question either exists, or it does not exist. If it exists, it either belongs to the ontology as described by the primitive notions, or it is derived from the ontology thus described. Hence, if one is committed to the existence of something without being able to derive it from the ontology as defined by the primitive notions, one has to enlarge the ontology so that it includes this thing as a further primitive.

The latter is at issue when it comes to the mind. If one has reservations about functional definitions in terms of a causal role for behaviour and shrinks back from going for elimination, then one has to endorse further primitives in the ontology when it comes to the mind. Thus, for instance, in the metaphysics that Chalmers (2012) proposes within the methodology of the Canberra plan, he endorses conscious experience as a further primitive beyond the physical ones. It is irrelevant here whether consciousness occurs only at certain places or times. If it exists and cannot be located in what is accepted as primitive, there is no other possibility but to endorse it as a further primitive in the ontology, however rare or abundant its occurrence in the universe may be.

This does not mean that consciousness (or whatever else may be endorsed as further primitive) constitutes a new level of reality with respect to a level of the world that is described by natural science. It just means that there are more primitives in the ontology than the ones admitted by natural science. Of course, one then has to spell out the relationship between these primitives. Employing the notion of levels of reality suggests that this work has been done, while in fact nothing in that respect has been achieved by employing this notion. The same goes for the notion of emergence: it suggests that something has been understood or even explained, while, in fact, no understanding or explanation has been provided. In particular, there is no point in seeking to avoid the debate about further primitives when it comes to the mind by employing a confused notion of emergence – that is, a notion that takes emergence to be opposed to reduction, but bases itself on the trivial sense of the emergence of new features at specific places and times in the universe. The confusion then lies in the suggestion that there can be the emergence of something within a naturalized, physicalist ontology without that something being located in the ontology as defined by the notions that are endorsed as primitive in a physicalist ontology.

4.3 Normative Functionalism and the Ontology of the Mind

There are two types of challenges when it comes to the mind: the challenge from conscious experience concerns features of which it is claimed that they do not admit of a functional definition (so called qualia). If this is so, they have to be accepted

as primitive: being intrinsic, qualitative features, there is no means available to locate them in an ontology in which they do not figure explicitly in the primitive notions that define the ontology. The challenge from rationality, by contrast, is of another type: there is no question of rationality consisting in qualitative, intrinsic features. The features characterizing rationality admit of functional definitions. But the challenge is that functional definitions in terms of roles for, finally, behaviour and thus the motion of matter are not the correct functional definitions when it comes to the mind, because they miss the normativity that characterizes rationality.

Indeed, the causal role functionalism that allows for the location of everything in the domain of the natural sciences in a primitive ontology of matter in motion is to be contrasted with a normative functionalism according to which the functional definition of mental concepts – insofar as these admit of a functional definition – is an affair of indicating their role in a normative network of justifications, that is, giving and asking for reasons. That normative network constitutes a realm of its own. It is related to behaviour through actions. But the point is that actions are not reducible to behaviour due to the normativity that they involve. Let us assume, at least for the sake of the argument, that this normative functionalism has a point and let us investigate its consequences for the ontology of the mind in the methodology given by the Canberra plan. In other words, let us consider what a dualism without levels can look like.

Normative functionalism was developed even earlier than the causal role functionalism that is standard today, namely by Sellars (1956) in his masterpiece "Empiricism and the philosophy of mind". Sellars (1956) is in the first place concerned with justification. He claims that (a) only something that has itself an epistemic status can justify something that has an epistemic status and that (b) nothing that is given to the mind has as such an epistemic status. The latter idea is what Sellars dismisses as the "myth of the given". Abandoning this myth implies that nothing that the mind of a person takes in from whatever external source can as such justify anything. Thus, for instance, sense impressions, construed as the effects of interactions of a person with the physical environment, cannot, qua being the result of physical *causal* processes, *justify* the beliefs of a person. By the same token, supposedly innate ideas – or ideas entering the mind through a causal relationship with God or a Platonic realm of ideas viz. Popper's world 3 –, cannot as such justify anything. The reason is that, with respect to whatever is given to her mind, the person has to take the attitude of endorsing what is given as a reliable source of knowledge in the circumstances at hand. Only thereby does she confer to it an epistemic status. Nothing comes as such with this status; it acquires this status by the way in which persons use it to form beliefs.

Taking something given as a reliable source of knowledge in the circumstances at hand is a holistic affair. It amounts to forming a belief that is linked up with other beliefs in such a way that the result is an overall coherent system of beliefs. Forming beliefs on the basis of what is given to the mind consists in navigating in what Sellars (1956) calls "the space of reasons". The system of beliefs is in continuous evolution, as new items enter that require adaptations within the system of beliefs to maintain its overall coherence. This system can therefore be related to what Quine (1951)

calls "the web of belief" and the procedure of adapting that web set out in his "Two dogmas of empiricism". Rejecting the myth of the given therefore leads to a holism of confirmation and justification in the guise of a coherence theory of knowledge, whereby coherence is the overall coherence with respect to the evidence received from external sources – in other words, the overall system that best explains this evidence.

Moreover, this is a social holism. When a person forms a belief – and be it a simple belief about everyday matters of fact –, she employs at least one concept. She thereby follows a rule that fixes what is correct and what is incorrect in applying the concept. In other words, the rule tells her how she *should* apply the concept. Furthermore, she follows a rule only if she is aware of her employing a concept being subject to a differentiation between correct and incorrect. This is what distinguishes rule-following from mere regularities of behaviour, and this the reason why beliefs are subject to a justification. Rule-following as necessary and sufficient condition for mastering concepts has been worked out notably by Wittgenstein in the *Philosophical Investigations* (1953, §§ 138–242) and the interpretation of Wittgenstein by Kripke (1982). Wittgenstein's argument is that only social interactions enable a person to distinguish between following a rule correctly and failing to do so. Only the interaction with others creates a distinction between what a person considers to be correct and what is correct in the eyes of others (see in particular Wittgenstein, 1953, § 202). That is why a social theory of meaning goes together with a normative theory of meaning (and *vice versa*): the view is that social, normative practices – and only they – determine meaning.

Brandom (1994, part one) spells this view out in terms of meaning being constituted by normative practices of commitment, entitlement and precluded entitlement. For instance, if under appropriate circumstances, a person utters the statement "The animal over there in the water is a whale", she thereby is committed to statements such as "The animal over there in the water is a mammal", she is entitled to statements such as "The animal over there in the water is huge" and she is precluded from being entitled to statements such as "The animal over there in the water is a fish". The meaning of the concept "whale" thus consists in the inferences that its use licences according to the norms of commitment, entitlement and precluded entitlement that are endorsed in a community. Accordingly, Sellars (1956, § 36) defines knowledge through its normative status:

> ... in characterizing an episode or a state as that of *knowing*, we are not giving an empirical description of that episode or state; we are placing it in the logical space of reasons, of justifying and being able to justify what one says.

In sum, the rejection of what Sellars (1956) denounces as the "myth of the given" leads to a justificatory, semantic and social holism in the guise of a social, normative theory of meaning.

Dismissing the myth of the given implies freedom of belief. Given the sensory input from the world – and whatever other input –, a person has to make up her mind as to what to believe. Kant already brought this point out by saying

> If an appearance is given to us, we are still completely free as to how we want to judge things from it. (Prolegomena § 13, note III; quoted from the translation Kant, 2002, p. 85)

This quotation implies that freedom including the free will of persons concerns not only actions, but also and already beliefs. A person has to make up her mind not only as far as her actions are concerned, but also as far as her beliefs are concerned, and be it beliefs about simple everyday matters of fact. She deliberates about beliefs in the same way as about actions.

The connection between freedom in belief and freedom in action is also brought out by McDowell when he describes what it would take for a wolf to entertain beliefs:

> A rational wolf would be able to let his mind roam over possibilities of behaviour other than what comes naturally to wolves. . . . [This] reflects a deep connection between reason and freedom: we cannot make sense of a creature's acquiring reason unless it has genuinely alternative possibilities of action, over which its thought can play. . . . An ability to conceptualize the world must include the ability to conceptualize the thinker's own place in the world; and to find the latter ability intelligible, we need to make room not only for conceptual states that aim to represent how the world anyway is, but also for conceptual states that issue in interventions directed towards making the world conform to their content. A possessor of *logos* cannot be just a knower, but must be an agent too; and we cannot make sense of *logos* as manifesting itself in agency without seeing it as selecting between options . . . This is to represent freedom of action as inextricably connected with a freedom that is essential to conceptual thought. (McDowell, 1995, § 3)

Freedom in belief thus goes together with freedom in action and *vice versa*. Failing to acknowledge either one of them would be an instance of falling victim to the myth of the given. Deliberation concerns beliefs in the same way as actions. As actions are not imposed on persons by given biological needs and desires, so beliefs are not imposed on them by given sense impressions. The question is "What should I believe?" in the same way as "What should I do?". With this freedom come in norms as the guides for beliefs and actions and thereby also justifications for the beliefs as well as the actions that a person adopts. That is why abandoning the myth of the given has a bearing on ontology: it brings out the freedom of persons both in employing concepts and in deciding how to act.

This freedom implies that persons cannot be located in the ontology of the natural domain. Any scientific theory including natural science as a whole – the scientific image in the terms of Sellars (1962) – is itself conceived, endorsed and justified in the normative web of giving and asking for reasons. When navigating in this web, a person has to presuppose the freedom to make up her mind about what to think and to do as primitive: any belief that she forms, any theory that she adopts is set up by her in exercising this freedom; taking it to be imposed on her from the outside would amount to falling back into the myth of the given.

Hence, one cannot claim that the matter in motion in the world imposes the theory that everything is matter in motion on us, because the theory itself is nothing but a configuration of the matter in motion in the sense that it is nothing beyond the beliefs that persons have, and these are realized by and thus identical with certain particle configurations in their brains. The reason is, again, that any such claim is

itself conceived, endorsed and justified in the normative web of giving and asking for reasons. Taking it to be imposed on us by the matter in motion in the world would be an instance of the myth of the given.

Rejecting the myth of the given thereby leads to an argument for persons being ontologically primitive: persons have to take decisions and thus to answer the question what they should do, including which beliefs and theories they should accept. Consequently, normativity is presupposed for the very formulation of a scientific theory. The referents of the theory – whatever the theory poses as existing in the world – cannot impose the acceptance of the theory on persons and justify it. In that sense – as the beings that formulate and justify theories in normative practices of giving and asking for reasons –, persons are primitive: whatever the theory is, persons have to conceive, endorse and justify the theory in question. Consequently, insofar as they formulate scientific theories and the scientific image as a whole, persons cannot be located or placed within what science poses as existing. One may go as far as to say that claiming that the scientific image includes persons as being located in its ontological primitives comes close to a performative contradiction: the content of the claim that everything is matter in motion contradicts its performance as *claim* that is situated in the normative web of giving and asking for reasons in which persons are primitive.

According to the method of metaphysics as set out in the quotation by Jackson at the beginning of this paper (the Canberra plan), there is a close link between epistemology and ontology: if, in the case at hand, the functional reduction of normative notions to the primitive physical notions fails, then not only have the normative notions to be recognized as irreducible and thus epistemologically primitive – that is, they have to be admitted as further primitive notions over and above the physical ones –, but also their referents have to be endorsed as ontological primitives, since they then cannot be located in the primitive physical entities. That is why epistemological irreducibility implies ontological irreducibility. In other words, there is no third way between either eliminating something or subscribing to an ontological commitment to it. On the Canberra plan, this either is a commitment to that something as ontologically primitive or comes with the obligation to establish how that something is located in what one admits as ontologically primitive by showing how its description can be reduced to a description in terms of the notions originally admitted as primitive. If such a reduction fails for principled reasons, both the notions in question and the entities they refer to have to be admitted as further primitives.

If persons as characterized by the normative attitudes that they adopt to one another are ontologically primitive, they can indeed be conceived in the same way as matter in motion on the proposal sketched out in the first section: both matter and persons are points that are structurally individuated through the relations in which they stand. Matter points are individuated by their position in a web of distance relations. Persons or mind points are individuated by their position in a normative web of rights and obligations, commitments, entitlements and precluded entitlements that concerns beliefs as well as actions. As all there is to the matter points are the distance relations in which they stand, so all there is to the mind points

are the normative relations into which persons enter through deliberating about what they should think and do. Hence, neither matter nor minds are characterized by any intrinsic features. The resulting view is a dualism, but not a dualism of any intrinsic features that distinguish minds from matter.

Both the distance relations and the normative relations are in continuous change. The normative relations change through every move that a person makes in her thoughts and actions. As the continuous change in the distance relations provides for an intertemporal identity of the matter points through the trajectories that they thereby trace out, so the continuous change in the normative relations provides for an intertemporal identity of the persons qua mind points.

The difference between matter points and persons or mind points lies in the difference in the relations that individuate them: distances that exist as a matter of fact versus norms that come into being through certain configurations of matter in motion adopting to themselves and others the attitude of taking themselves and the others to be situated in a web of rights and obligations. In adopting such an attitude, certain particle configurations create themselves as persons: in doing so – and only in doing so – are they persons. This difference in the relations implies that the normative relations only exist as long as persons continue to exist by adopting these attitudes. More precisely, the distance relations that characterize and individuate material objects are accessible from a third person perspective, that is, the point of view from nowhere and nowhen that characterizes science. They exist as a matter of fact independently of whether or not anyone conceptualizes them. By contrast, the normative relations that individuate persons qua mind points are accessible only from within participating in the practices that determine them, as pointed out, for instance, by Sellars (1962, section VII). This follows from the characterization of being a person through adopting a normative attitude towards oneself and others: to access the norms that are determined by these attitudes, one has to adopt this attitude towards the beings in question and thereby to participate in the normative practices in question, thus contributing to shaping these norms.

That notwithstanding, there are sufficient physical conditions for persons to come into being. The ability to engage in social, normative practices is located in and thus identical with certain particle configurations. One can formulate a biological explanation of this ability in terms of the enhancement of fitness that cooperation between humans provides (see, for instance, Tomasello, 2014). Nonetheless, once these practices come into being, the norms that are determined in them are not located in the sphere of facts. They are not further facts in the world. They exist, as the matter in motion exists; but they are accessible only from within participating in these practices and thereby contributing to shape them. There is no perspective from nowhere and nowhere available to access these practices.

Hence, the difference between persons and matter in motion, between mind points and matter points, is not one in existence or truth conditions. Existence and truth are unequivocal. Either something exists or it does not exist. Either a proposition is true, or it is not true. The difference is one of accessibility: without contributing to shape them in the case of taking note of facts in contrast to accessing

norms only by contributing to determine what they are in adopting the attitude of treating oneself and others as persons.

Consequently, we face the problem of how to bring science and what characterizes us as persons together not because our perspective or our knowledge is somehow limited. We can formulate scientific theories that apply to the universe as a whole from a perspective of nowhere and nowhen. Cosmology always did so and continues to do so. These theories (or some successors of them) may be true. The point at issue is that any theory, including a theory of the universe as a whole construed from the point of view of nowhere and nowhen, can be formulated only from within participating in social, normative practices that determine its content. There is no other possibility for a theory or a whole image of the world, whatever its content may be, to be conceived, endorsed and justified. This, then, amounts to an argument to the conclusion that insofar as persons formulate theories, they are ontologically primitive: they cannot be located in anything else that a theory poses as primitive, for posing that something presupposes persons as those beings who conceptualize, endorse and justify the theory in question in their practices of giving and asking for reasons.

However, there is no question here of levels of reality. Quite to the contrary, one blurs the distinction between matter in motion and persons if one talks in terms of different levels of reality on which these are situated. There are not different levels of matters of fact, properties, or objects. Both matter and persons have to be endorsed as primitive, as the method of metaphysics demanded by the Canberra plan brings out (if indeed persons can neither be eliminated nor located in an ontology defined by the primitive notions that characterize matter in motion). But the difference between them is not a difference between levels of reality; it is the difference between facts and norms.

Acknowledgements I'd like to thank the editors for the invitation to contribute to this volume. The paper has been improved by comments from Erez Firt for which I'm grateful.

References

Bell, J. S. (1987). *Speakable and unspeakable in quantum mechanics*. Cambridge University Press.
Brandom, R. B. (1994). *Making it explicit. Reasoning, representing, and discursive commitment*. Press.
Chalmers, D. (1996). *The conscious mind. In search of a fundamental theory*. Oxford University Press.
Chalmers, D. (2012). *Constructing the world*. Oxford University Press.
Correia, F., & Schnieder, B. (Eds.). (2014). *Metaphysical grounding: understanding the structure of reality*. Cambridge University Press.
Esfeld, M. (2020). *Science and human freedom*. Palgrave Macmillan.
Esfeld, M., & Deckert, D.-A. (2017). *A minimalist ontology of the natural world*. Routledge.
Esfeld, M., & Sachse, C. (2011). *Conservative reductionism*. Routledge.
Hall, N. (2009). *Humean reductionism about laws of nature*. Unpublished manuscript. http://philpapers.org/rec/HALHRA

Hemmo, M., & Shenker, O. (2015). The emergence of macroscopic regularity. *Mind & Society, 14*, 221–244.

Jackson, F. (1994). Armchair metaphysics. In J. O'Leary-Hawthorne & M. Michael (Eds.), *Philosophy in mind* (pp. 23–42). Kluwer.

Kant, I. (2002). *The Cambridge edition of the works of Immanuel Kant. Volume 3. Theoretical philosophy after 1781. Edited by Henry Allison and Peter Heath*. Cambridge University Press.

Kim, J. (1998). *Mind in a physical world. An essay on the mind-body problem and mental causation*. MIT Press.

Kripke, S. A. (1982). *Wittgenstein on rules and private language*. Blackwell.

Leibniz, G. W. (1890). *Die philosophischen Schriften von G. W. Leibniz. Band 7. Edited by C. I. Gerhardt*. Weidmannsche Verlagsbuchhandlung.

Lewis, D. (1966). An argument for the identity theory. *Journal of Philosophy, 63*, 17–25.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy, 67*, 427–446.

Lewis, D. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy, 50*, 249–258.

Lewis, D. (1986). *Philosophical papers. Volume 2*. Press.

Mach, E. (1919). *The science of mechanics: a critical and historical account of its development. Fourth edition. Translated by Thomas J. McCormack*. Open Court.

Maudlin, T. (2019). *Philosophy of physics. Quantum theory*. Princeton University Press.

McDowell, J. (1995). Two sorts of naturalism. In R. Hursthouse, G. Lawrence, & W. Quinn (Eds.), *Virtues and reasons: Philippa Foot and moral theory* (pp. 149–179). Oxford University Press.

Price, H. (2004). Naturalism without representationalism. In M. de Caro & D. Macarthur (Eds.), *Naturalism in question* (pp. 71–88). Harvard University Press.

Quine, W. V. O. (1951). Two dogmas of empiricism. *Philosophical Review, 60*, 20–43.

Russell, B. (1912). On the notion of cause. *Proceedings of the Aristotelian Society, 13*, 1–26.

Sellars, W. (1956). Empiricism and the philosophy of mind. In H. Feigl & M. Scriven (Eds.), *The foundations of science and the concepts of psychology and psychoanalysis* (pp. 253–329). University of Minnesota Press.

Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.), *Frontiers of science and philosophy* (pp. 35–78). University of Pittsburgh Press.

Shenker, O. (2017). Flat physicalism: Some implications. *Iyyun. The Jerusalem Philosophical Quarterly, 66*, 1–15.

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

Wittgenstein, L. (1953). *Philosophical investigations. Translated by G. E. M. Anscombe*. Blackwell.

Chapter 5
Can the Flat Physicalist Tell Us What a Physical Entity Is?

Erez Firt

Abstract Physicalism is the thesis that everything is physical, or that everything supervenes on the physical. It is unique as a metaphysical doctrine in that it has a close relation to the physical sciences, expressed by the physicalist claim that what exists in the world, i.e., the physical, is what physics maintains exists. This theory-based conception troubled Carl Hempel (Synthese, 45, 139–199, 1980), who in response formulated what later came to be known as Hempel's Dilemma (HD). In this paper we will examine an extension to a particular reply to HD known as the current-physics reply (CPR), which is an attempt to maneuver between the two horns of HD by relying on a new, improved version of physics that is close enough to current physics so as not to be a vague future version. It is straightforward empirical scientific approach that attempts to avoid metaphysical implications and thereby dismiss the problems raised by HD. Although boldly refreshing, I argue that this approach fails to avoid the horns of HD.

5.1 Introduction

Physicalism is the thesis that everything is physical, or that everything supervenes on the physical. Physicalism is unique as a metaphysical doctrine in that it is closely related to the physical sciences. This relationship is expressed by the physicalist claim that what exists in the world, i.e., the physical, is what physics maintains exists. This theory-based conception[1] troubled Carl Hempel (1969, 1980), who in response formulated what came to be known as Hempel's Dilemma[2] (HD).

[1] See Stoljar (2021), §4.1.

[2] Following Crane and Mellor (1990), who referred to the reduction of different scientific fields to physics and famously described the problem as a dilemma: "We must first ask to what physics the

E. Firt (✉)
University of Haifa, Haifa, Israel

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_5

For physicalists who look to physics to determine what is physical, this dilemma presents a key problem with number of facets: If physical principles are based on current physics, it follows that they are incomplete and are almost surely, at least partially, false. If physical principles are based on a future version of physics, they are inherently vague as this version of physics does not yet exist. And if physical principles are based on an ideal complete future version of physics, we run the risk of trivializing physicalism. In all of the above cases, HD implies that our conception of the physical is not clear enough to serve as the foundation for such a prominent metaphysical view.

HD has inspired a good number of replies from physicalists. In this paper we examine an extension to a particular reply to HD known as the current-physics reply (CPR). This extension is not minor adjustment of CPR, but an attempt to maneuver between the two horns of HD by relying on a new, improved version of physics that is close enough to current physics so as not to be a vague future version. One interesting version of this extended reply is offered by Hemmo and Shenker (2019), and is embodied in their physicalist identity theory of "flat physicalism."[3] This version of the extended current-physics reply, henceforth ECPR, is a straightforward empirical–scientific approach that attempts to avoid metaphysical implications and thereby dismiss the problems raised by HD. In that it differs from other notable recent attempts to respond to the challenge of HD, specifically the via negativa approach and the attitudinal view.[4]

This paper is structured as follows: In Sect. 5.2 we outline the main proposals or directions of responses to HD in the literature. Section 5.3 presents the physicalist theory of flat physicalism (FP) as part of the ECPR approach. Section 5.4 raises a deep, inherent problem that is left unanswered by this extended reply in general: Whether the ECPR proponent's research is a success or a failure, all outcomes of an ECPR approach leave them entangled in the horns of HD. Section 5.5 concludes the paper.

5.2 A Brief Survey of Some Proposed Solutions to Hempel's Dilemma

Supporters of physicalism proposed various ways to either overcome HD or live with it. Let us describe in outline and very briefly the main proposals or directions of responses to the Dilemma in the literature, highlighting the CPR-related views. For

RIP [Reduction in Principle] principle is supposed to be applied: to present physics, or to some hypothetical future physics? This question poses a dilemma." (p. 188) See also Buzaglo (2017) and Bokulich (2011) for comprehensive reviews of Hemple's dilemma.

[3] See Hemmo and Shenker (2015, 2019, 2021a, b) and Shenker (2017).

[4] These approaches and views are outlined in the next section.

convenience, the main proposals in the literature can be divided into the following major groups.

One group of proposals claims that it is reasonable to think that the ontology is well described by contemporary physics, at least approximately (after all, the success of contemporary physics is a salient motivation for accepting physicalism in the first place). According to these views, physics will not change much, so they reject the claim that we don't know the nature of future physics, thus rejecting the claim of Horn 2. A number of views within this group also reject the claim of Horn 1, believing that current physics does explain the phenomena satisfactorily, and to the extent that future physics will significantly change, these changes will not affect the part of physics that is relevant to those explanations. This subgroup represents what is sometimes known as the current-physics reply, henceforth CPR (see e.g., Smart (1978), Lewis (1994), and Bokulich (2011)). By contrast, Melnyk (1997, 2003) who also rejects Horn 1 of the Dilemma makes no claim that current physics is even approximately true; instead, he argues that current physics is more likely to be true relative to its rivals, where an as yet unformulated future theory doesn't count as a relevant rival.

A second group rejects Horn 2 of the Dilemma and argues that referring to some future physics, that is yet unknown, is the right move, and it is neither trivial nor empty or vague. Among the adherents of this view, Poland (1994), for example, conjectures that the integration of the mental into its future-physical place will be carried out in a similar manner to the incorporation of the electromagnetic theory into fundamental physics, and therefore although we don't know anything about future physics, we do know that current physics will find the place in which the mental can be integrated into it in this sense.

A third proposal to characterize "physics," called via negativa, was put forward in the context of psycho-physical reductionism, and characterizes the "physical" as non-mental, thus avoiding the need to address the details of the science of physics, either present or future (see Montero (1999)). These views are based on the idea that physics is causally complete (or closed), roughly that physical events are preceded by physical causes (see Spurrett and Papineau (1999); criticism by Gillett and Witmer (2001), and a rebuttal by Montero and Papineau (2005)).

A fourth approach attempts to find definite characteristics of physics, such as, the fact that physics describes the most fundamental elements of the universe, and perhaps also the fact that physics is the only kind of scientific theory in which the laws are strict and have no ill-defined exceptions in the sense that also includes genuine probabilistic laws (see Dowell (2006) for an approach in this direction). However, it is unclear that such an approach can avoid the addition of something along the lines of the via negativa approach in order to block the possibility that facts about the mental are incorporated into physics as primitives (see Wilson (2006)).

A fifth group tries to avoid the Dilemma altogether by denying that physicalism is a sentence with a truth value (or an expression of a belief about the world that may be true or false); thus in a sense embracing the Hempelian conclusion and its close relative, the so-called Chomsky's challenge (see Chomsky (1968, 2000, 2003); Poland (2003)) that physicalism is vacuous because the concept of the physical lacks

content. Instead, this approach advances the idea that physicalism means adopting an attitude or a stance to form one's theory of the world according to what the best available theories of physics at the time say exists. Supporters of this view in different versions are Hellman (1985); van Fraassen (2002); Poland (2003); and Ney (2008).[5]

There is also a more general approach to HD (see Firt, Hemmo and Shenker (2022)), which claims that the dilemma is neither solved nor solvable; and nevertheless it is not a threat to physicalism per se. The reason for this is that HD can be generalized in such a way that its domain of application is much wider: it applies to all the theories that explain the phenomena or high-level facts by appealing to some underlying deep structure, and that use methodologies that allow for a change of these explanations. Physicalism indeed offers a changeable deep structure account of the high-level facts, and is therefore subject to HD. According to the authors, mind-body dualism, for example, also offers changeable deep structure accounts and is therefore no less subject to HD.

5.3 An Extended Current-Physics Reply

The Current-Physics Reply to HD holds that in the physicalist view physics is nearly complete. It follows that any future adjustments to physics will be minor and most probably irrelevant to the metaphysics of the mental, thus posing no threat to physicalism. Smart (1978) specifically addresses the mind–body question by stating that "whatever revolutionary changes occur in physics there will be no important lesson for the mind–body problem." Lewis (1994) is less specific: "We may reasonably think that present-day physics already goes a long way toward a complete and correct inventory [of all the fundamental properties and relations that occur in the world]." As is Bokulich (2011): "[We have] good scientific reasons for believing that the future development of physics will be irrelevant for the metaphysics of the mental, the biological, the sociological and other common terrestrial phenomena."

In this paper, however, I argue against the version of ECPR suggested by Hemmo and Shenker[6] as part of their understanding of identity theories. ECPR proposes a middle way to navigate between the horns of HD—not relying on current physics, yet not relying upon some vague future version of physics. ECPR suggests that a future version of physics that solves current physics problems will provide explanations for all high-level theories of the special sciences based on fundamental physical theories. When discussing the problems of "current physics," ECPR proponents refer to the problems of what proponents of CPR call "ordinary physics." Thus, ECPR's close relationship to CPR is expressed in the assumption

[5] See Stojar (2021, Sec. 4) for additional problems and responses.

[6] See footnote #3.

that only minor changes will be needed to arrive at a future version of physics that will be stable. As mentioned earlier, flat physicalism (FP) is an identity theory that embraces the principles of ECPR.[7] In general, FP refers to physicalism as a scientific hypothesis rather than a metaphysical stance. Proponents of the FP view take statistical mechanics (SM) and thermodynamics as a paradigm for the way fundamental physical theories should account for all higher order theories in physics and the special sciences.[8] The authors refer to SM as the theory of physical kinds[9] and attempt to apply the way SM explains thermodynamics to all other cases:

> The account of the thermodynamic regularities by statistical mechanics proceeds in two major steps. The first step consists in associating each thermodynamic quantity (such as temperature, pressure, volume) with a mechanical quantity (e.g., temperature is a function of molecular motion, such as average kinetic energy). Each mechanical quantity here is a macrovariable or function of the mechanical microstate; it is an aspect of the microstate given by its partial description, and therefore it gives only partial information about the microstate of the system . . . The second step in accounting for the thermodynamic regularities in statistical mechanics consists in recovering the laws of thermodynamics, in particular the approach to equilibrium and the second law of thermodynamics, from the mechanical laws governing these mechanical macrovariables. (ibid: 462-3)

The authors employ SM notions of microstate[10] and macrovariable[11] to construct an identity theory which is *flat*, i.e., does not require additional metaphysical relations, such as *emergence*,[12] *supervenience*,[13] *realization*,[14] *grounding*,[15] etc., between the hierarchical levels of ontology. In fact, in FP "there are no levels of reality . . . [I]nstead of high level and low level, what we have are different aspects, given by different descriptions, of the state of the universe . . . [I]f there is only one level of reality and of description, there is no room, and no need, for discussing inter-level relations." (Shenker, 2017: 4) For example, according to FP the following is a flat identity statement: the volume of a gas (a thermodynamic magnitude) *is* the distribution of positions of the gas particles (a mechanical macrovariable consisting of the set of possible microstates, associated with this volume).[16] The flat physicalist has in mind the generalization of these kinds of identity statements in her attempt

[7] However, supporters of FP are not bound to the conclusion that their view dismisses HD.

[8] See Shenker (2017, §2).

[9] See Hemmo and Shenker (2019: 461) and Shenker (2017: 5–7).

[10] The term microstate is used in SM to denote the complete mechanical state of the system of interest, or of the world.

[11] Macrovariables are sets of microstates that share a certain aspect (which reflects partial information about the microstate of the system). In other words, when we observe a certain aspect of a system, the system may be in any of the microstates belonging to the set of microstates associated with this specific aspect or macrovariable.

[12] See O'Connor and Wong (2015).

[13] See McLaughlin and Bennett (2011).

[14] See McLaughlin and Bennett (2011, §3.6).

[15] See Bliss and Trogdon (2014).

[16] For the full explanation, See Hemmo and Shenker (2019, §2).

to construct an identity theory that will eventually explain higher order theories by employing only fundamental physical theories.

Our flat physicalist is a theoretical physicist (or a philosopher of physics) who seeks to address and explain the problems of "ordinary and terrestrial"[17] current physics: e.g. open questions in quantum field theory, the measurement problem, the ontological status of the wave function, arrow of time, etc., henceforth present-day physical problems. Once the physicalist formulates a version of physics that addresses these open issues, her work is done. At this point, according to Hemmo and Shenker, our flat physicalist's project can either succeed or fail. If she succeeds, her version of physics should explain everything, i.e., fundamental physical theories will provide explanations for all high-level theories of the special sciences.

Thus ECPR, as adopted by FP, leads to the following assumptions:

1. *Solvability*: The main assumption is that a future version of physics that solves the ordinary, common, terrestrial problems of current physics will account for all non-physical theories.
2. *Stability*: This future version of physics will be a stable version in that it will not change drastically and will at most require fine tuning.
3. *Closeness*: This future version of physics will require only minor changes to the part of current physics that proponents of ECPR regard as "ordinary." Radical changes may occur in cosmology or quantum gravity, to give just two examples, but as far as brain science and the physics relevant to the study of consciousness are concerned, proponents of ECPR follow proponents of CPR in assuming that only minor inessential changes will take place.

5.4 Some Problems for ECPR

Before addressing the problems of ECPR, let us address the problem that proponents of CPR and ECPR do manage to avoid, i.e. the assumption of a complete true final version of physics. What is a "complete true final version of physics" and why is it problematic? Several philosophers have attempted to address the concept of a future complete version of physics: Armstrong (1991: 186) for example, writes about the set of properties the physicist will appeal to in the end. Others are more explicit: Loewer (1996: 103) says that "fundamental physical properties . . . are the properties expressed by simple predicates of true comprehensive fundamental physical theory," and Horgan (1994: 472) goes further to say humans are constituted of entities that are "the kind posited in (an ideally completed) physics". The notion of true and completed physics is problematic because it requires the fairly controversial

[17] I hereby refer to the descriptions used by supporters of CPR, as regards the physics currently available to us: Bokulich (2011) referred to "common terrestrial phenomena" as something current physics explains very well, and Lewis (1994) referred to "ordinary matter under mild conditions", as something that is well understood.

metaphysical assumption that we can actually reach the truth about the physical world and not just approach it in some sense. This is a far-reaching assumption, for even the notion of approximate truth in the context of scientific realism is a bone of contention in the realist–anti realist debate.[18] A more crucial problem in this context is that assuming a complete true physics makes physicalism trivially true. "For what is a true and completed physics, save for one that accounts for the fundamental nature of everything?" (Montero, 2005:179) For example, as far as the mind–brain debate is concerned, a completed physics will by definition account for the mental as well, thus making physicalism, at least as a theory of mind, trivially true.

A proponent of ECPR quite rightly rejects the idea of a final complete true theory of physics and embarks on a scientific project to reach a near-future version of physics. This version would solve the problems faced by "ordinary" current physics, and once reached would comply with *stability* and *closeness*. The project of reaching this version of physics can succeed, or fail. And in the case of failure, there are two other options: the ECPR-ist can either admit that she was wrong, or continue with her version of physicalism after all.

If the above-mentioned scientific project succeeds, i.e., physicists of the future succeed in developing a version of physics that can solve "ordinary" physical problems that puzzle present-day physicists, the flat physicalist believes that all non-physical theories will be explained by these future physical theories in terms of identities of higher-level magnitudes with macrovariables. According to Hemmo and Shenker, in the case of the mental an ECPR may not be needed at all, for it is possible that current physics—in particular quantum field theory (QED)—will suffice, as it may be applied to explain the workings of the physical brain, which in turn should explain the mental. Bokulich (2011) in his concluding words concurs with this claim: "In the case of brain processes, we have compelling reasons to believe that the most important processes are chemical and electrical, and that all of this is safely within the domain of QED." (ibid: 650). To be sure, quantum field theory or a near-future version of physics that will solve present-day physical problems, may lead to robust scientific knowledge regarding the inner workings of brain processes. For physicalists such knowledge will lead to a physical explanation of the mental. However, this close relation to CPR has disadvantages in the form of several common objections.[19]

Let us examine the option of a future relatively stable version of physics that solves present-day physical problems. If it is complete (as *stability* may suggest) then we fall back to the idea of a final complete true version of physics, and this is not what a proponent of ECPR has in mind. Thus, this version solves present-day physical problems and explains all non-physical higher theories, but is incomplete and hence has unsolved problems and uncovered areas of reality. The flat physicalist's hope is that solving present-day physical problems will lead to

[18] See Chakravartty (2017, §3.4).

[19] See Crook and Gillett (2001), Pineda (2006), and Buzaglo (2017) who explains and extends their objections. See also Stoljar (2010). I will not rehearse these objections here.

a version of physics that complies with *stability* and *closeness,* and hence is not exposed to the second horn of Hempel's dilemma. History teaches us that past scientists[20] who held the same beliefs were proven wrong, and there is nothing distinctive about current physics and its problems that can support or justify the flat physicalist's belief. Moreover, the problems faced by present-day physicists are fundamental and significant, and the identity relationship required by flat physicalists is strict and demanding. In other words, to accomplish such goals physicists would need to make radical changes in current physics, thus violating *closeness* and in turn possibly raising other currently unpredictable problems and thus violating *stability* as well. Additionally, we should also note that the search for solutions for what ECPR proponents consider the remote problems of physics ("unordinary physics," e.g. cosmology, quantum gravity, theory of everything) may raise unpredictable problems that might also affect "ordinary" physics.

In the case of FP, the physicalist may reply that she is a scientist who took upon herself an empirical project of formulating a physical theory that solves present-day physical problems. She is not a metaphysician interested in grandiose metaphysical claims regarding the nature of reality. Her interests lie solely in the success of her future physical theory, i.e., its acceptance by the scientific community, its empirical success, etc. This reply is in accordance with the initial statements of the flat physicalist, but it does not suffice, for FP cannot be viewed as a mere scientific theory. Conceived by philosophers, FP is associated with a family of metaphysical theories (identity theories of physicalism) and formulated using metaphysical concepts such as identity relations, reduction, and more. But even if we grant the flat physicalist that her project is a scientific one carried out by physicists who have no interest in metaphysical questions, then their product is still a version of physics, appealed to by (metaphysical) physicalists who are vulnerable to the horns of Hempel's dilemma.

In case the above-mentioned scientific project fails, there will be a non-empty set of non-physical theories that cannot be accounted for by fundamental physical theories. At this point, the flat physicalist will face the following two options:

1. Admit that the physicalist approach is incorrect and declare FP to be false.
2. Readjust her goals in hope that future versions of physics will be able to do the work that was supposed to be done by the version of physics that solves all present-day physical problems.

Option (a) is clear so let us take a closer look at option (b). Any proponent of ECPR who accepts this option clings to a dogmatic hope that even though there are still non-physical theories that cannot be accounted for by fundamental physical theories, future versions of physics will eventually succeed in explaining them.

[20] At the end of the nineteenth century and the beginning of the 20th there was an "odd sense of completion", as Steven Weinberg calls it, among scientists. Lord Kelvin, Albert Michelson, Philipp von Jolly (Planck's teacher) are just a few of the scientists who were quoted as saying (roughly) that the fundamental laws of physical science have already been discovered and what remains is precise measurement.

Hence, despite the failure, the dogmatic physicalist believes that ECPR is still the answer. Note, however, that each such cycle (i.e. failure and readjustment of goals) weakens *closeness* (for each such cycle pushes us further and further from current physics) and proves *stability* wrong (for there will be new emerging problems to be solved).

This dogmatic hope stems from physics enjoying a special privileged status in the eyes of physicalists, namely that it will no doubt reach a future state able to explain ontological claims of non-physical theories. This reads as scientism, i.e., the idea that for any sort of intellectual inquiry to be rational and acceptable, it must conform to the models of science, or as Hacker (2007: 3) states: "[Scientism is the] *illicit* extension of the methods and forms of explanation of the natural sciences." Some extreme supporters of scientism would even claim that the natural sciences are the only source of real knowledge. In our case, and perhaps in every case of physicalism, scientism is manifested in the belief that whatever non-physical theories remain unaccounted for by the best version of physics currently available to us, they will no doubt be explained by future versions of physics. Physicalists who hold such a position can never be proven wrong.

To sum up the difficulties with basic ECPR assumptions presented at the end of Sect. 5.2:

Closeness asserts that no significant changes should be made to "ordinary" current physics in order to reach the flat physicalist's future version of physics. This is in the spirit of ECPR, as it is an extended, closely-related version of CPR. However, this also implies that "ordinary" current-physics is stable, as no substantial amendments are needed to solve present-day physical problems. This reveals the similarity between CPR and ECPR, making ECPR vulnerable to the objections mentioned above.[21] Thus, *stability* and *closeness* together imply that both current physics and the future version that solves present-day physical problems are stable, at least as far as the "ordinary" core is concerned, which is not far from suggesting that we have reached the end of "ordinary" physics. This is a problematic assumption.[22] Without these assumptions, as mentioned above, arriving at the flat physicalist future version of physics might require radical changes to current physics, which will no doubt raise unpredictable new problems and expose this version to the first horn of HD.

[21] See footnote #19.

[22] The construal of the first horn of HD with a reference to the historical failures of scientific theories – also mentioned in Buzaglo (2017) – can be easily linked to what is called the "meta-pessimistic induction" argument against scientific realism. As Buzaglo (2017) stresses, there are at least four different construals of the first horn of HD, therefore I am reluctant to commit myself to this specific one.

5.5 Concluding Remarks

To conclude, in this paper we examine an extended version of the current-physics reply to Hempel's Dilemma. A proponent of ECPR, represented here by the flat physicalist for purposes of clarity and illustration, puts forth the scientific hypothesis that everything is physical—physical being what the physics that will solve present-day physical problems will maintain exists. This version of physics, the ECPR proponent hypothesizes, will fundamentally explain (in the manner discussed in Sect. 5.3) all other high-level non-physical theories. I argue that this puts the proponent of ECPR at the starting point of one of three possible routes: If correct, Hempel's dilemma still looms over the future version of physics. If incorrect and her hypothesis is false, then depending on the ECPR proponent's acceptance of this failure, she is either dogmatic or, well, just wrong. In addition, other difficulties are exposed upon a closer examination of the basic assumptions of ECPR.

In summation, my argument with the ECPR proponent can be summed up as follows: It is agreed by both sides that her main assumption, *solvability*, may require radical changes in current physics. The ECPR-ist hopes that such changes will not affect "ordinary" physics, thus sustaining *stability* and *closeness*. However, there is in actuality no basis for such hope and a mistake in this case will make her thesis vulnerable to HD. On top of that, as mentioned above, other changes in more remote parts of physics might also affect "ordinary" physics, thus again entangling ECPR in the horns of Hempel's Dilemma.

References

Armstrong, D. (1991). The causal theory of mind. In D. M. Rosenthal (Ed.), *The nature of mind* (pp. 181–188). Oxford University Press.

Bliss, R., & Trogdon, K. (2014). Metaphysical Grounding. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/grounding/

Bokulich, P. (2011). Hempel's Dilemma and domains of physics. *Analysis, 71*(4), 646–651.

Buzaglo, D. (2017, Unpublished). *Hempel's Dilemma and the formulation of physicalism.*

Chakravartty, A. (2017). Scientific Realism. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/scientific-realism/

Chomsky, N. (1968). *Language and mind*. Harcourt Brace and world.

Chomsky, N. (2000). *New horizons in the study of language and mind*. Cambridge University Press.

Chomsky, N. (2003). Replies. In L. Antony & N. Hornstein (Eds.), *Chomsky and his critics*. Blackwell.

Crane, T., & Mellor, H. D. (1990). There is no question of physicalism. *Mind, 99*, 185–206.

Crook, S., & Gillett, C. (2001). Why physic alone cannot define the 'physical': Materialism, metaphysics and the formulation of physicalism. *Canadian Journal of Philosophy, 31*(3), 333–360.

Dowell, J. L. (2006). Formulating physicalism. *Philosophical Studies, 131/1*. Special Issue.

Firt, E., Hemmo, M., & Shenker, O. (2022). Hempel's Dilemma: Not only for physicalism. *International Studies in Philosophy of Science*. https://doi.org/10.1080/02698595.2022.2041969

Gillett, C., & Witmer, G. D. (2001). A physical need: Physicalism and the via negativa. *Analysis, 61*, 302–308.

Hacker, P. M. S. (2007). Wittgenstein and the autonomy of humanistic understanding. *E-Journal Philosophie der Psychologie*, *9*.
Hellman, G. (1985). Determination and logical truth. *The Journal of Philosophy, 82*(11), 607–616. https://doi.org/10.2307/2026415
Hemmo, M., & Shenker, O. (2015). The emergence of macroscopic regularity. *Mind and Society, 14*(2), 221–244.
Hemmo, M., & Shenker, O. (2019). Two kinds of high-level probability. *The Monist, 102*(4), 458–477.
Hemmo, M., & Shenker, O. (2021a). Flat physicalism. *Theoria* (.), forthcoming.
Hemmo, M., & Shenker, O. (2021b). *Why functionalism is token-dualism*. this volume.
Hempel, C. (1969). Reduction: Ontological and linguistic facets. In S. Morgenbesser et al. (Eds.), *Essays in Honor of Ernest Nagel*. St Martin's Press.
Hempel, C. (1980). Comments on Goodman's ways of worldmaking. *Synthese, 45*, 139–199.
Horgan, T. (1994). Physicalism. In S. Gluttenplan (Ed.), *A Companion to philosophy of mind* (pp. 471–479). Blackwell Publishers.
Lewis, D. (1994). Reduction of mind. In S. Guttenplan (Ed.), *A companion to philosophy of mind*. Blackwell Publishers.
Loewer, B. (1996). Humean supervenience. *Philosophical Topics, 24*(1), 101–127.
McLaughlin, B., & Bennett, K. (2011). Supervenience. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/supervenience/
Melnyk, A. (1997). How to keep the 'physical' in physicalism. *Journal of Philosophy, 94*, 622–637.
Melnyk, A. (2003). *A physicalist manifesto: Thoroughly modern materialism*. Cambridge University Press.
Montero, B. G. (1999). The body problem. *Noûs, 33*(2), 183–200.
Montero, B. G. (2005). What is the physical? In B. McLaughlin & A. Beckermann (Eds.), *Oxford handbook of the philosophy of mind* (pp. 173–188). Oxford University Press.
Montero, B., & Papineau, D. (2005). A defense of the Via Negativa Argument for Physicalism. *Analysis, 65*(3), 233–237.
Ney, A. (2008). Physicalism as an attitude. *Philosophical Studies, 138*, 1–15.
O'Connor, T., & Wong, H. Y. (2015). Emergent properties. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/properties-emergent/
Pineda, D. (2006). A mereological characterization of physicalism. *International Studies in the Philosophy of Science, 20*, 243–266.
Poland, J. (1994). *Physicalism*. Clarendon.
Poland, J. (2003). Chomsky's challenge to physicalism. In L. Antony & N. Hornstein (Eds.), *Chomsky and his critics*. Blackwell.
Shenker, O. (2017). Flat physicalism: Some implications. *Iyyun, the Jerusalem Philosophical Quarterly, 66*, 1–15.
Smart, J. J. C. (1978). The content of physicalism. *The Philosophical Quarterly, 28*, 339–341.
Spurrett, D., & Papineau, D. (1999). A note on the completeness of 'physics'. *Analysis, 59/1*, 25–29.
Stoljar, D. (2010). *Physicalism*. Routledge.
Stoljar, D. (2021). Physicalism. *Stanford Encyclopedia of Philosophy*. https://plato.stanford.edu/entries/physicalism/
van Fraassen, B. C. (2002). *The empirical stance*. Yale University.
Wilson, J. (2006). On characterizing the physical. *Philosophical Studies, 131*, 61–99.

Chapter 6
How Context Can Determine the Identity of Physical Computation

Nir Fresco

Abstract Computational explanations in the cognitive sciences span multiple levels of analysis. The indeterminacy of computation complicates the endeavour of answering the question 'What does a particular neural—or physical—system do?' in *computational* terms. A single physical process may often be described equally well as computing several different mathematical functions—none of which is explanatorily privileged. But at which level of analysis is the computational identity of a physical system *P* fixed? Some argue that the computational identity of *P* is wholly exhausted by *P*'s functional or narrow physical structure. Others argue that *contextual* factors also play a role in determining *P*'s computational identity, but they diverge on *what* that role is precisely. Yet others argue that contextual factors *essentially* determine the identity of *P*. This chapter surveys some of these views and ultimately claims that the environment *can* and often *does* play an important role in fixing the computational identity of *P*, thereby proposing a new, long-arm functional strategy for individuating computation.

6.1 Introduction

Computational explanations in the cognitive sciences span multiple levels of analysis—from a detailed biophysical model of single neurones, through a neurocognitive description of neural activity between different brain regions, to the identification of algorithmic models of some cognitive phenomena and the specification of the mathematical function computed by some brain circuits. The identification of *the* mathematical function being computed by a physical system—be that a brain circuit, or a single neurone—may be complicated by the fact that

The work presented in this chapter is partly based on Fresco (2021).

N. Fresco (✉)
Ben-Gurion University of the Negev, Be'er Sheva, Israel

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_6

Table 6.1 Electrical gate *G*'s output falls within the ranges shown in the output column when its two inputs fall within the ranges shown in the input columns

Input-channel 1	Input-channel 2	Output-channel
1–3 V	1–3 V	1–3 V
1–3 V	4–6 V	1–3 V
1–3 V	1–3 V	1–3 V
4–6 V	4–6 V	4–6 V

some such functions have *other* "isomorphic copies"—a term that will shortly be clarified using a simple example. Thus, any computational description that is based on identifying such mathematical functions as part of explaining the explanandum may be based on indeterminate computation.

If that is true, then computational explananda in the cognitive sciences are likewise susceptible to exhibiting this phenomenon. Neurocognitive explanations regularly confront the question: 'What does a specific neural structure do and how does it do it?'. Answers that hypothesise that the structure concerned computes some specific mathematical function, which has an isomorphic copy, should arbitrate between the possible competing functions. At least prima facie, neither mathematical function is epistemically privileged as a description of the structure's physical behaviour. We turn next to describe a simple physical system whose behaviour is multiply-specifiable using two *Boolean* functions, yet this phenomenon is not limited to Boolean functions.

Consider a simple, electrical Boolean gate *G* with two input-channels and a single output-channel. *G*'s physical behaviour is described in Table 6.1. If the voltage range 1–3 V represents False and 4–6 V represents True, Table 6.1 turns out to be the standard truth-table for Boolean conjunction. Thus, *G* computes conjunction (or is an AND-gate). However, if the voltage range 1–3 V represents True and 4–6 V represents False, *G* computes an isomorphic copy of conjunction, namely: inclusive disjunction (or is an OR-gate). (Conjunction and disjunction are considered *dual* functions in Boolean logic.) The gate's computing either conjunction or disjunction illustrates the indeterminacy of computation. A similar moral applies to an electric gate that computes either NAND or NOR function and a gate that computes either XOR or XNOR function (and to other dual Boolean functions).

There seems to be agreement that it is necessary to identify relevant features that determine what computation a given physical system *P* actually performs (Fresco et al., 2021). In philosophy of computation, whether or not a given account of computation is able to settle the indeterminacy of computation when it arises has been deemed a litmus test for the adequacy of that account. However, what those features are and the level at which they are invoked remain open questions. Bishop (2009), Sprevak (2010), and Shagrir (2020), for example, appeal to features at a *semantic* level of analysis that render it determinate what computation is performed by *P*. Proponents of the mechanistic view of computation diverge on the precise level of analysis. Dewhurst, for example, argues that the relevant features that render it determinate what *P* computes are identified at a narrow *physical* level of analysis (2018). Coelho-Mollo argues that the computational identity of *P* is determined at

a functional level of analysis (2017). Piccinini advocates a wide, short-arm view about determining the computational identity of *P*.

As a backdrop for proposing a new, long-arm functional individuative strategy—as a middle way between the wide, short-arm strategy and the semantic one, the main contenders in the debate are examined. The main motivation for the proposed strategy is that some interesting cases of indeterminacy in biological systems cannot always be settled by appealing to activities that do not exceed beyond the relevant sensory receptors and motor neurones of the system—as the short-arm functional individuative strategy suggests. Real things in the world may count as the inputs and outputs of the computation performed (Fresco, 2021). However, such long-arm strategy need not collapse into the semantic individuative strategy as is briefly argued in Sect. 6.5.

The claim defended in this chapter is that the system-environment interaction often plays an important role in fixing *P*'s computational identity, thereby advocating the long-arm functional strategy. The system-environment interaction can be used to settle which of two mathematical functions $f(x)$ or $g(x)$ *P actually* computes. Harbecke and Shagrir (2019) have likewise recently argued that contextual factors essentially determine the computational identity of *P*. Although some mechanists partially agree with this claim (e.g., Miłkowski, 2017; Piccinini, 2015; Coelho Mollo, 2019), they diverge on *what* this role is. This chapter will examine this disagreement, and argue for a new, long-arm functional strategy according to which *P*'s inputs and outputs may, in some cases, also be realised *outside* the system itself.

The chapter is organised as follows: Section 6.2 discusses two influential mechanistic strategies of individuating computation very narrowly. In Sect. 6.3, we examine the semantic individuative strategy, according to which at least in *some* interesting cases of computational indeterminacy semantic constraints are needed to determine the system's computational identity. Section 6.4 discusses a pluralistic view encompassing both mechanistic and semantic individuative strategies relative to different explanatory contexts. Section 6.5 advances a long-arm functional individuative strategy as midway between the wide, short-arm mechanistic view and the semantic view of computational individuation. Section 6.6 briefly responds to an explanatory challenge to the long-arm functional individuative strategy.

6.2 Computational Individuation at Narrow Physical and Functional Levels of Analysis

In this section, we examine two different mechanistic positions concerning computational individuation. In a nutshell, a "mechanism for a phenomenon consists of entities and activities organised in such a way that they are responsible for the phenomenon" (Illari & Williamson, 2012, p. 120). The mechanistic explanatory strategy, then, is to decompose the explanandum into its spatiotemporal constituent parts, and to discover how their causal interactions and structural relations are

responsible for producing (or maintaining) the explanandum. To see how different proponents of the mechanistic view of computation diverge on computational individuation, let us first discern various relevant levels of analysis.

Which level of analysis is apt for computational individuation, and how can it avoid (or settle) potential indeterminacies? Table 6.1 may be said to specify the behaviour of some physical system at a purely *physical* level of analysis: in response to voltages in specific ranges, the system produces voltages in a specific range. This provides one systematic specification of the system's physical behaviour. Once the voltage ranges are mapped onto logical True (or 1) and False (or 0), the corresponding table provides a logical specification of the system's behaviour (describing either conjunction or disjunction in our case). Call this the *logical* level of analysis. The physical or logical level may turn into a *semantic* level of analysis, if the variables are mapped onto some content, such as numbers. Semantic content, however, may, at least in principle, be assigned arbitrarily to the corresponding variables (even if the assignment is systematic), hence, infinitely many different semantic specifications of the system's behaviour are possible. Of course, theories of semantic ascription would typically be very constrained and non-arbitrary. Computational mechanists, such as Dewhurst, Coelho Mollo and Piccinini, disagree about the level of analysis that is apt for computational individuation.

6.2.1 *Dewhurst's Individuative Strategy: Losing Computational Equivalence*

Let us start from Dewhurst's position that proposes to individuate computation at a purely physical level (2018). As a computational mechanist, Dewhurst can specify the computational identity of a given system in virtue of three key ingredients (Fresco & Miłkowski, 2019). The first is digits: what are the unique digits processed by the system and how many are there? (The answer is '2' in relation to gate *G* above: [1–3 V] is one digit, and [4–6 V] is the second, distinct digit.) The second ingredient is the processing unit(s) that operates on these digits: how many processing units are in use by the given system? (The answer is '1' in relation to *G*. But a more complex computing system that comprises many Boolean circuits may have many distinct processing units.) The third—important—ingredient is the input-output relations in which the digits partake in the encompassing system. (Table 6.1 specifies *G*'s input-output relations.) By Dewhurst's lights, the computation performed by *G* (i.e., a function from physical inputs to physical outputs) can be fully spelled out in terms of these three ingredients.

If that is right, then computational states and processes can be individuated without invoking any logical or semantic content. The motivation for such an individuative strategy becomes apparent once we realise that the type of indeterminacy described above does not manifest itself: it occurs *only* once we specify the system's behaviour at the *logical* level. *G*—as it is described by Table 6.1—indeterminately

Table 6.2 Hydraulic gate *H*'s inputs and corresponding outputs specified in water pressure measured in Litre per Second (LpS)

Input-channel 1	Input-channel 2	Output-channel
0.1–0.5 LpS	0.1–0.5 LpS	0.1–0.5 LpS
0.1–0.5 LpS	1–1.5 LpS	0.1–0.5 LpS
1–1.5 LpS	0.1–0.5 LpS	0.1–0.5 LpS
1–1.5 LpS	1–1.5 LpS	1–1.5 LpS

computes conjunction *or* disjunction depending on how the variables (voltage ranges) are mapped onto True and False. But *if* Table 6.1 (plus 'digits' and 'processing units') provides all the theoretical posits necessary for individuating *G*'s computation, then at least prima facie—as Occam's razor dictates—this individuative strategy seems appealing: computation is individuated by non-semantic transformations of digits, and indeterminacy is thereby avoided.

But looks can be deceiving: the economic efficiency of this strategy comes at a cost, namely giving up computational equivalence and multiple realisability.[1] Computational equivalence is a central idea in computer science: two physical systems may compute the same function even if the physical magnitudes they operate on are different. One system may traffic in voltages, the other in *different* voltages or even fluid pressure, and they may still both compute conjunction. (Of course, such computational equivalence occurs at the logical level.) Thus, the hydraulic gate *H* described in Table 6.2 may be computationally equivalent to *G*. If both the low voltage range and, likewise, the low water pressure are mapped onto False, *G* and *H* compute conjunction (and *mutatis mutandis* they may both compute disjunction). Since at a purely physical level, Tables 6.1 and 6.2 describe distinct physical behaviours of the respective systems, Dewhurst's strategy individuates them as *different* computations.

But the problem cuts even deeper. Because "the physical structure of two computing mechanisms is always going to be distinct, and it is unclear whether we can draw any non-arbitrary boundary between the structures that are relevant or irrelevant to computational individuation" (Dewhurst, 2018, p. 110). Any two conventional AND-gates in one's smartphone—made of the same materials, based on the same blueprint, by the same manufacturer—turn out to be *computationally distinct* (since any minute difference in their voltage ranges is enough). The idea of computational equivalence is, thus, lost on Dewhurst's strategy.

The closely related idea of multiple realisability is likewise threatened by this individuative strategy.[2] Cognitive explananda that are multiply realisable bestow an explanatory edge to *causes* over physical *constituents*. The physical constituents

[1] Note, however, that the computational mechanist need not bite the bullet and pay this price (Fresco & Miłkowski, 2019).

[2] Some computational mechanists indeed deny that multiple realisability is an essential feature of physical computation. Miłkowski, for one, claims that "there are no facts of the matter that could easily establish that a given computational capacity is actually multiply realized" (2016, pp. 29–30). The computational mechanist faces a dilemma. Either computational explanations are functional, and, thus, cannot fully explain the structural aspects of mechanisms, or they provide

in one realisation that constitute a given cognitive phenomenon (e.g., visual object identification) will not necessarily be the constitutive elements in another realisation of the same phenomenon. "There is little reason to believe that cognitive and neural entities and activities must be similarly organized. In complex systems, what looks stable and robust at one scale may not be so at another scale" (Stinson, 2016, p. 1603). Insofar as *distinct* neural structures can give rise to the *same* cognitive function by computing a *specific* mathematical function, the mathematical computation may be more stable as a cause than the particular constituents. The present computational individuative strategy is incompatible with the common and compelling explanation of multiple realisability of cognitive functions in terms of computational functions.

In sum, for those who think that computational equivalence and multiple realisation should be preserved as important principles in the computational sciences, including cognitive science, "the physical level is [simply] too fine-grained" (Coelho Mollo, 2017, p. 3493) and so the present individuative strategy fails to deliver the goods.

6.2.2 *Coelho Mollo's Individuative Strategy: Moving to a Functional Level*

Realising that this is too great a price to pay, Coelho Mollo extends Dewhurst's strategy and fixes computational individuation at a *functional* level of analysis "in which the only structural considerations at play are having appropriate degrees of freedom" (Coelho Mollo, 2017, p. 3494). By classifying *computational* phenomena as a proper subset of *teleofunctional* phenomena, Coelho Mollo's individuative strategy gains an important explanatory advantage; it draws a boundary between computing and non-computing systems. Planetary motions, hurricanes, and tides are, thus, excluded as non-computational phenomena.

How does Coelho Mollo's individuative strategy work? In essence, it draws on the principle of 'equivalence classes'—a technical notion in logic. In this strategy, however, "[e]quivalence classes are defined by input values that lead to uniform behaviour of the whole device—the differences in value to which the device is sensitive and which are uniformly transformed into new values" (ibid). To see how this definition works, consider another gate, *G*,* which is very similar to *G* (described by Table 6.1), but whose voltage ranges are (2–4 V) and (5–7 V) instead. *G** and G are computationally equivalent. Why? Because both *G* and *G** respond to two distinct equivalence classes of acceptable physical inputs (voltages), and yield the same equivalence classes of physical outputs (voltages) in response. Each such equivalence class is a *digit*. In *G*, the first equivalence class (or digit) is (1–3 V) and

full structural detail, but give up multiple realisability as an essential feature of computation (Haimovici, 2013, p. 178). Miłkowski, like Dewhurst, opts for the second horn of the dilemma.

in *G** it is (2–4 V); *G*'s second equivalence class is (4–6 V), and *G**'s is (5–7 V). Thus, to some extent computational equivalence is preserved on Coelho Mollo's individuative strategy.

Another important explanatory advantage of this strategy is that computational equivalence also holds between systems of a different physical makeup. The hydraulic gate *H* above (described by Table 6.2) is computationally equivalent to *G* and *G** at the functional level. Because what matters to computational individuation is the overall functional profile that defines these three gates. *H* shares the same functional profile of *G* and *G**, since it is sensitive to, and responds uniformly and in the same way to the same number of equivalence classes. Whilst the *physical descriptions* of *G*, *G**, and *H* are clearly *distinct*, at the *functional* level, their description is *identical*. Whether the equivalence classes are based on voltages or water pressure is, supposedly, irrelevant for computational individuation. The functional level of analysis, presumably, exists as an intermediary between the physical and logical levels.

Nevertheless, this individuative strategy raises two main worries. The first one concerns the notion of an equivalence class. In logic, an equivalence relation—over a given set A—is one that is reflexive, symmetric, and transitive.[3] That is, it satisfies specific logical properties. An equivalence relation divides A into equivalence classes based on these properties. The equivalence relation in Coelho Mollo's strategy, though, is not characterised as rigorously. 1.42 V and 2.41 V, for example, are classified as belonging to the same equivalence class with respect to *G*. Because *G*'s behaviour is sensitive to, and responds in the same way to both values. The equivalence relation is one that partitions the set [(1–3 V) ∪ (4–6 V)] into the two corresponding ranges. In that sense, each being an equivalence class is a trivial matter: they are defined as such (on the basis of *G*'s systematic physical behaviour to be sure!). But which properties should 1.42 V and 2.41 V (and *every other* possible voltage value in the relevant range) satisfy to belong to the *same* equivalence class? One might further object that the partitioning method is *ex post facto*: first, identify the *logical function(s)* that the physical gate computes, and, then, fix the equivalence relation on that basis. How else can we determine which functional differences "are not relevant to the regimented input–output transformations of equivalence classes of physical states across the system" (Coelho Mollo, 2017, p. 3496)?

Coelho Mollo's reply would likely be that the equivalence relation is determined by the behaviour to be explained (2017, p. 3490). First, the system should be functionally decomposed in light of the behaviour to be explained (e.g., producing a specific output when only two inputs are above a certain threshold). Second, the functional component that systematically produces a *specific* output for these *specific* ranges of inputs (e.g., taking two inputs and behaving differentially when they are received) should be identified. This observed differential behaviour partitions

[3] Suppose that '∼' is a binary equivalence relation on A. Reflexivity means that for all a ∈A, a ∼ a. Symmetry means that for all a, b ∈A, if a ∼ b, then b ∼ a. And transitivity means that for all a, b, c ∈A, if a ∼ b and b ∼ c, then a ∼ c.

the system into two equivalence classes, if they have two different functional roles. This discovery method, however, partitions the system's behaviour into equivalence classes only *ex post facto*: producing a specific output when only two inputs are above a certain threshold, for example, is simply another way of describing conjunction.

A second worry is that this strategy seems to divorce computational *individuation* from computational *explanation* (Shagrir, 2020, p. 4098). A cognitive capacity that is explained by means of specifying the mathematical or logical function that a mechanism computes typically qualifies as a *computational* explanation. (Coelho Mollo would describe it as a *mathematical model* explanation, though.) A discovery that the locust's visual neurone computes *multiplication* in order to trigger an escape response to a looming object at just the right time (Gabbiani et al., 2002) *is* explanatory of that capacity (under the relevant theoretical and empirical constraints). Coelho Mollo, however, claims that "logical individuation is at least one step above computational individuation" (2017, p. 3495). It follows, then, that computational individuation is separate from computational explanation, even when the latter is couched in purely formal terms.

Why is that problematic? Coelho Mollo need *not* deny, of course, that describing the locust's visual neurone as computing multiplication is *explanatory*. Rather, he would claim that computational individuation—at the functional level—is what allows us to make sense of the mathematical model explanation in terms of the neurone's multiplication. Thus, Coelho Mollo has to reject the common claim that explanatory practices in the cognitive sciences are roughly aligned with Marr's tri-level explanatory hierarchy (e.g., Anderson, 2015; Blokpoel, 2018; Hardcastle & Hardcastle, 2015)—as others have recently suggested (e.g., Bickle, 2015; Love, 2015). Computational indeterminacy arises precisely at Marr's top-level that specifies the problem solved by the system in terms of input-output relations. And identifying the relevant input-output relations is an important step in figuring out *why* the system does what it does and *how*. Input-output (I/O) equivalence does not entail functional equivalence. Any two I/O equivalent algorithms may go through different sequences of states intermediate between these inputs and outputs. Hence, it is "pertinent to [further] inquire as to which state(s) the system occupies in the process of producing its output(s) for some given input(s)" (Buller, 1993, p. 158). And, indeed, scientists have to figure out which algorithm is likely used to compute that function and propose a plausible biophysical model that supports their hypothesis (Jones & Gabbiani, 2012). How does Coelho Mollo's individuative strategy fit with these computational practices?[4]

[4] A possible answer is that this individuative strategy somehow specifies the system's algorithmic level. The "different functional profiles [of two computing systems] would [result a difference] in their capacity to carry out logical and mathematical functions" (Coelho Mollo, 2017, n. 20) (3495, fn 20). Evaluating this answer, though, exceeds the scope of this chapter.

In sum, the present individuative strategy certainly fares better than Dewhurst in preserving a narrow version of computational equivalence by simply giving up on some implementational details at a purely physical level. It remains unclear, however, how a *functional* equivalence relation may be further regimented in an analogous manner to its *logical* counterpart. Moreover, adopting this strategy comes at the cost of being at odds with at least some explanatory practices in computational cognitive science.

6.3 Computational Individuation at a Semantic Level of Analysis

Having examined two *narrow* individuative strategies, we now turn to the *wide* counterpart: the semantic individuative strategy. According to this strategy, representation is necessary for computation. A strong version of the semantic strategy may require that *only* semantic properties figure in the computational individuation of a state (or process). However, both Sprevak (2010), and Shagrir (2001, 2020) advance a weaker, and thus more plausible, individuative strategy, according to which a computational state (or process) is partially individuated by semantic properties and partially by non-semantic properties. Thus, Shagrir, for example, accepts that the relation of implementing an automaton by a physical system need not be individuated semantically, but claims that computational individuation proper does require semantic individuation (2020, p. 4088). Hence, he argues that *G* (and *H* for that matter) simultaneously implement both the AND and OR automata. To determine the computational identity of such systems, on the present view, semantic constraints are required.

Relatedly, Sprevak argues that I/O equivalence is a necessary, but not a sufficient, condition for computational identity (2010, p. 269). The "respective inputs and outputs [. . . of our *G* and *H* gates] are different, [. . .] so different as to not have any physical or functional properties in common" (2010, p. 268). Sprevak asks what their I/O equivalence may consist in; Coelho Mollo's response is 'similarity in equivalence classes and their respective degrees of freedom'. The former, however, claims that the I/O equivalence of *G* and *H* consists in their respective inputs and outputs *representing* the same thing. Nevertheless, the *type* of representational content is *unconstrained*; it may be mathematical, proximal, distal, narrow, or wide. Even if the inputs and outputs of an AND-gate are labelled with the numeral '0' or '1', such syntactic labelling is *still* representational content. Similarly, "[n]o physical, structural, or functional property decides", so Sprevak claims (2010, p. 269), whether *G* computes conjunction or disjunction. It is a difference in representational content.

One of the main arguments in support of the semantic individuative strategy concerns the semantic individuation of tasks. Lee (2021) summarises it very nicely; let us call it the 'task individuation' argument.

P1: Computations feature in explanations of task performance.
P2: Tasks are individuated semantically.
C: Hence, computation requires semantic individuation.

P1 may seem uncontentious at first, and, thus, may be accepted by computational mechanists. A closer inspection reveals that some proponents of mechanistic explanations may reject P1 as irrelevant to the case in point (Miłkowski, personal communication). For tasks, as such, are not the phenomena to be explained, but are often only experimental *effects*. And "the phenomena we typically call 'effects' are incidental to the primary *explananda* of psychology" (Cummins, 2000, p. 140), namely cognitive capacities (e.g., learning capacity, the capacity for depth perception, and planning capacity). Given that mechanists would argue that scientific explanation is *phenomenon*-based, and task performance is only secondary, P1 should be rejected.[5]

Computational semanticists, such as Sprevak and Shagrir, however, endorse both P1 and P2. There are at least two good reasons for endorsing P2. The first one has just been discussed: computational I/O equivalence between systems like *G* and *H* can be easily defended (thereby, indirectly, also supporting the idea of multiple realisability). The second reason is fending off Putnam- and Searle-like triviality arguments according to which every (complex enough) physical system computes every Turing-computable function (and there are infinitely many such functions!).

As said above, the computational mechanist denies any appeal to semantic properties for computational individuation. Unlike the narrow individuative strategies proposed by Dewhurst and Coelho Mollo's, Piccinini accepts that *contextual* factors can play a role in determining the computational identity of a physical system. However, in response to the task argument, he claims that "a (non-semantic) functional individuation of computational states is sufficient to determine which task is being performed by a mechanism, and hence which computation is explanatory in a context" (Piccinini, 2015, p. 43).

Shagrir agrees that the computational indeterminacy exhibited by *G*, for example, can, indeed, be settled by appealing to non-semantic functional properties. Without exceeding the boundaries of the encompassing system that contains *G*, one might appeal, say, to arm movement as described by Table 6.3 below. If, as the table shows, the connected arm *only* moves when both inputs are within the high voltage range, then *G* may be said to compute conjunction (rather than disjunction). No semantic property needs to be invoked to determine *G*'s computational identity in this case.

To show that Piccinini's individuative strategy cannot deal with more intricate cases of computational indeterminacy,[6] Shagrir proposes a simple, yet clever,

[5] Piccinini also adds that the task individuation argument would not go through, if we rejected the assumption that explanatia and their explananda must be individuated by the same properties (2015, p. 40).

[6] It should be stressed here that Shagrir's example of the tri-stable system exhibits a different kind of indeterminacy (resulting from how different microstates are grouped together) from the one discussed in relation to gates *G*, *G**, and *H* above (resulting from how state types are labelled). An

Table 6.3 An electrical gate *G* with arm movement that is triggered only when both inputs are within the high voltage range

Input-channel 1	Input-channel 2	Output-channel	Arm movement
1–3 V	1–3 V	1–3 V	✗
1–3 V	4–6 V	1–3 V	✗
1–3 V	1–3 V	1–3 V	✗
4–6 V	4–6 V	4–6 V	√

Table 6.4 An electrical gate similar to *G* with the original low voltage range divided in two, and three types of corresponding arm movement: no movement (between 0° and 45°), medium movement (between 45° and 90°), and high movement (greater than 90°)

Input-channel 1	Input-channel 2	Output-channel	Arm movement
1–2 V	1–2 V	1–2 V	None (e.g., 0–45°)
1–2 V	2–3 V	2–3 V	Medium (e.g., 45–90°)
2–3 V	1–2 V	2–3 V	Medium (e.g., 45–90°)
1–2 V	4–6 V	2–3 V	Medium (e.g., 45–90°)
4–6 V	1–2 V	2–3 V	Medium (e.g., 45–90°)
2–3 V	2–3 V	2–3 V	Medium (e.g., 45–90°)
2–3 V	4–6 V	2–3 V	Medium (e.g., 45–90°)
4–6 V	2–3 V	2–3 V	Medium (e.g., 45–90°)
4–6 V	4–6 V	4–6 V	High (e.g., >90°)

modification of the voltage ranges on which *G* operates. The resulting gate (see Table 6.4 above) now has *three* voltage ranges instead of just *two*; it is a tri-stable system. Suppose that the low voltage level is now (1–2 V) and (2–3 V). (Thus, grouping them both together still gives us [1–3 V] as before.) Shagrir argues that this construction enables us to individuate movement as either (a) high movement only (i.e., when both inputs are within the high voltage range) or (b) medium movement *plus* high movement (i.e., in all possible input combinations except for {[1–2 V], [1–2 V]}). If we adopt the first option, the gate computes conjunction, but if we adopt the second, the gate computes disjunction. How can any of the mechanistic individuative strategies decide which functional kinds are relevant in identifying the computation that is actually performed? This challenge has remained unanswered.[7]

analysis of the relation and difference between these two kinds of indeterminacy exceeds the scope of this chapter and is undertaken elsewhere (Papayannopoulos et al., 2022).

[7] Piccinini (2020, pp. 153–154) asserts that such cases of indeterminacy are addressed differently in natural and artificial computing systems. In the case of natural systems, we should identify (a) the capacity of interest, (b) the structures that fulfill that capacity, and (c) the specific organisation that enables those structures to fulfill the capacity. However, the tri-stable system described by Table 6.4 is an artificial one, and, thus, we "define the correct equivalence classes between [its] microstates as we please" (ibid). The *specific* mathematical function that is computed by that system depends, then, on a *choice* made by the engineer who designed and built the system.

The upshot of the task individuation argument and this last example is that at least in some explanatory contexts, the functional and semantic tasks are not co-extensive. The last example illustrates that "arm movements by themselves do not suffice to determine the units of the computation, and hence, the computation itself" (Harbecke & Shagrir, 2019). Thus, sometimes even "the system plus its immediate causal environment are not [...] sufficient for fixing the actual computations performed by the system" (ibid). The computational semanticist, thus, concludes that in order to determine the computational identity of at least some physical systems semantic constraints *are* required.

6.4 Computational Individuation Along Multiple Levels of Analysis

Having examined both the mechanistic and semantic individuative strategies, we now turn to briefly discuss Lee's pluralistic view of computational individuation with respect to these two strategies (2021). This view is based on there being multiple levels of analysis that pertain to computational properties and relations. Different properties and relations are relevant to the scientific categorisation practices depending on the explanatory context and the computational explanandum. Computational individuation along these levels can, thus, inform different contextual explanations relative to specific epistemic interests without thereby entailing anti-realism about computation (Lee, 2021, p. 241).

Lee's pluralistic view pertains to *three* hierarchical levels of analysis, but it seems that a *fourth* one should be added. The first—narrow functional—level in his hierarchy concerns the intrinsic functional properties of the computing system in question, *S*. It essentially corresponds to Coelho Mollo's individuative strategy. The next one—wide (short-arm) functional level—pertains to the role of *S* in some higher-level, encompassing mechanism, thereby considering S' interaction with its immediate context via its input and output channels. This level essentially corresponds to Piccinini's individuative strategy, which is further elaborated in Sect. 6.5. The third—semantic—level corresponds to the semantic individuative strategy and it concerns the relations between *S* and the distal states of affairs *S* represents.

According to this view, each level further constrains the one below it. Properties and relations specified at the narrow functional level constrain the possible computations that *S* may perform by having suitable degrees of freedom. At the wide (short-arm) functional level, properties and relations outside *S* are also considered, thereby further constraining the possible computations that *S* may perform. At the semantic level, relations between *S* and some relevant distal states of affairs are included when the sensitivities of the wider causal nexus are insufficient for determining S' computational identity. In cognitive science, semantic considerations place constraints on computational models, thereby helping to model the internal structure of the mechanisms contributing to the performance of a given cognitive

task (Miłkowski, 2017). Computational individuation, then, may occur at *any* one of these three levels.

It seems, though, that taking Dewhurst's individuative strategy into account requires adding another level: a *narrow physical* level. Lee claims that "Dewhurst's thesis [is] that computation is individuated at the level of narrow functional properties" (2021, p. 235), but, as noted above, this individuative strategy is based on non-functional transformations of digits. For that reason, Tables 6.1 and 6.2 above describe distinct *narrow physical behaviours* and, thus, denote *different* computations. On Coelho Mollo's individuative strategy, on the other hand, these two tables describe the *same* computation, since they are deemed *functionally* equivalent (in terms of the underlying equivalence classes). Thus, the properties and relations at the narrow functional level further constrain those specified at the narrow physical level. Moreover, if the intrinsic functional properties of the computing system "closely correspond to the notion of computation explicated in computability theory" (Lee, 2021, p. 222), then Dewhurst's individuative strategy cannot correspond to the narrow functional level. Because, to reprise, adopting this strategy results in abandoning the computational equivalence principle. A pluralistic view that accommodates both the mechanistic and semantic strategies should, so it seems, encompass four—or, rather, *five* as will be conditionally argued below—levels.

The upshot of this pluralistic view is that each level may provide equally legitimate descriptions of the computing system concerned, even though a higher level constrains the one below it. The reason that computational individuation at each level may be equally legitimate is based on a proposed separation between ontological and epistemic considerations (Lee, 2021, p. 236). Ontologically, a computing system may be self-contained or part of a larger encompassing system (e.g., a Boolean gate leaving the production line or one that is embedded in the motherboard chipset). Epistemically, a scientist or an engineer may be interested in her explanation to analyse the system either in isolation or as part of the wider context. Narrow physical or narrow functional properties may be *sufficient* for individuating computation in some specific manner; but wide functional (or semantic) properties may be *necessary* in another explanatory context (ibid). Neither *individuative route* is ontologically privileged, on this view. Nevertheless, given that the levels are hierarchically organised, it seems that a higher level is at least more *epistemically* privileged. For a higher level *also* pays attention to those properties and relations manifested at the lower levels.

By way of concluding this section, we note that this hierarchy of levels is basically metaphorical or heuristic. There exists no purely 'physical' or 'functional' level. Even if lower levels are considered more detailed than higher ones (higher levels abstract away from less relevant particulars, such as the energetic footprint of a physical computing process), there may be less detail in some *physical* model of a given phenomenon than a psychological, and thus *semantic*, model. We may (be misled to) think that a narrow physical level is supposedly the most elementary level of analysis that encompasses all possible physical elementary detail (i.e., the physical microstructure). But, then, the voltages specified in Tables 6.1, 6.3 and 6.4

are certainly not *elementary* physical properties. The elementary properties (e.g., quantum fields) might be those implied by the quantum mechanical equations that specify the behaviour of the computing system concerned. In that case, it is not clear, for example, how Dewhurst individuative strategy would work. The notion of level is notoriously fraught with difficulties (see, e.g., Brooks & Eronen, 2018) and should, hence, be handled with care.

We next turn to defend a long-arm functional strategy for individuating computation as midway between the mechanistic wide functional view and the semantic view. Arguably, this strategy may possibly correspond to yet another level in Lee's pluralistic view about computational individuation.

6.5 Long-Arm Functional Individuation of Computation

To position the proposed long-arm functional strategy in Lee's pluralistic hierarchy, let us focus on Piccinini's individuative strategy—which is short-arm, yet wide—and briefly on Shagrir's prolepsis of the long-arm strategy. Piccinini's strategy is short-arm, because the computational inputs and outputs have to be realised *within* the system itself. It is wide, though, since the computation concerned does not supervene only on the internal states and/or properties of the system itself. The long-arm strategy, on the other hand, adopts an ecological approach to computation (Wells, 1998): some computational inputs or outputs may be realised *outside* the system itself.

6.5.1 Wide, Short-Arm Individuation of Computation

Piccinini claims that understanding the nature of wide individuation requires an epistemic distinction between functionally relevant and irrelevant properties of the physical computing system (Piccinini, 2015, pp. 139–140). Drawing this distinction, on his view, requires knowledge of (a) which of the system's properties are relevant to its computational inputs and outputs, and (b) how they are relevant to the computational explanandum. This knowledge, in turn, requires an understanding of the way(s) that the system interacts—via inputs and outputs—with the context in which it is embedded.[8] This may lead us to conclude that Piccinini, in fact, concedes that at least in some cases, computational individuation is by *wide* content. If that were so, the mechanistic individuative strategy would supposedly collapse into its semantic counterpart.

[8] Coelho Mollo would similarly argue that the functional decomposition of the computing system depends on a capacity of interest, and this capacity may often be determined in part by the context in which it is embedded.

Piccinini, unsurprisingly, denies this consequence on the basis of two reasons (2015, pp. 140–141). First, he argues that the relevant functional properties are not *very* wide: they concern the interaction between the system and its immediate context via the system's input and output transducers. In artificial computing systems, the boundaries may be drawn at the forces exerted on input devices, such as trackpads, and the outputs produced by output devices, such as the screen monitor or printer. In biological systems, the wideness of the relevant properties required for computational individuation "does not even reach into the organisms' environment; it only reaches sensory receptors and muscle fibers" (ibid).

The second reason is that the mechanistically relevant properties are to be identified under the empirical constraints set by the suitable natural science or engineering methods. The computational identity of the system concerned may be discovered and individuated without appealing to any semantic properties. Thus, we are left with only short-arm factors for computational individuation.

However, a computational mechanist need not limit herself only to short-arm factors. Understanding how mechanisms, including computational ones, *actually* function often requires to situate them in their operational context. Bechtel's citation nicely captures this idea.

> The behavior of mechanisms is highly dependent on conditions in their environments, including any regularities that occur there. But these are not discovered by looking inside the mechanism to the parts and operations or how these are organized. They must be discovered by examining the environment in which the mechanism operates and employing tools appropriate for such inquiry. (Bechtel, 2009, p. 559)

On the other side of the spectrum, Shagrir advances a full-blown externalist individuative strategy. He claims that (a) whilst wide, short-arm individuation can indeed eliminate *some* cases of computational indeterminacy, *others remain* (see Table 6.4 and the discussion in Sect. 6.3), and, hence, (b) one possible route is "going even more external, to the outside environment" (Shagrir, 2020, p. 4102). Shagrir rightly requires, however, that a functional strategy that extends all the way into the environment and resolves all cases of indeterminacy should be shown to be (a) plausible, and (b) preferable to a semantic individuation of computation. In what follows, let us modestly take up only the first requirement; the second must await another opportunity.

6.5.2 A Functional Long-Arm Individuative Strategy

According to the long-arm functional strategy, computation is understood ecologically: as encompassing *both* the computing system and its surrounding environment (cf. Wells, 1998). Let us unpack this characterisation and defend the plausibility of this strategy using a toy, but realistic, example of a shared physical subsystem S in rodents.

Suppose that *S* receives two inputs: one from the hypothalamus in the form of the orexin hormone (which is involved in the sleep and wake cycle as well as energy balance), and another from the visual system. Orexin signals hunger when the organism's blood glucose levels are low, prompting the organism to search for food. The visual input, specifying the contours and textures of a visual object, signals the presence of an object, which *is likely to* be edible, within a visible distance from the organism. *S* produces a single output signal that is sent to the motor cortex, and when this output signal exceeds a certain threshold, it functions as a seek-food command. Such a computational description of *S* is clearly mechanistic. But absent further constraints *S* might still be indeterminate between an AND- and an OR-characterisation.

Why is that? Depending on the specific organism and its interaction with the environment, it may be the case that only when both inputs are "positive" (i.e., both the orexin input and the visual input exceed a certain threshold), *S* sends a seek-food command as a "positive" output. But if one of the inputs (or both) is "negative" (i.e., the relevant input threshold is not reached), then the output signal does not exceed the relevant threshold, and so no seek-food command is sent. This description is consistent with conjunction. However, it is likewise plausible that organisms of another related species would seek food even when one input is "positive". That is, if either blood glucose levels are low (thereby secreting orexin above a certain threshold) or a target object is within sight, the organism may forage for food (in response to S' seek-food motor output). In this scenario, the computational description of *S* is consistent with (inclusive) disjunction.[9] Roughly the same *S* can be used to compute two different mathematical functions (more on the 'roughly' qualification in footnote 10).

The specific organism-environment interaction can play a role in fixing S' computational identity. The computational identity of *S* can be fixed by S' biological function (cf. Coelho Mollo, 2019). The contextual factors that are relevant to determining S' biological function, however, may extend beyond the organism's sensory receptors and muscle fibers—as in Piccinini's wide, short-arm strategy. For that reason, "wide mechanistic explanations can [and should] be used by all researchers interested in the interaction of cognitive systems with their environments" (Miłkowski et al., 2018).

Suppose that *S* is a subsystem in the *hopping mouse*. The foraging behaviour of any species depends on the location and consumption of available resources, securing and storing these resources, existing competition with conspecifics and other species, and the risk of predation. It is quite plausible that, on average, a positive energy budget by the hopping mouse is expected only when both inputs to *S* are "positive". *S* may likely perform conjunction in the hopping mouse.

[9] An *exclusive* disjunction (XOR) interpretation under these circumstances is implausible. For it entails that when the organism is hungry and sees food, it does not reach out to grab it.

Another rodent, however, such as the *golden hamster*, might exhibit a different behaviour, if it were equipped with a similar subsystem.[10] Why? Because the amount of food hoarded by this organism increases significantly when food becomes available after being in short supply. Nevertheless, the amount of food this hamster typically consumes remains unchanged from pre-fast levels (Buckley et al., 2007). Such behaviour does nothing to decrease the hamster's appetite, and will likely continue until the food stored in its cheek pouches is actually chewed and swallowed, thereby resulting in an increased blood glucose level, and a decrease in the secretion of orexin. The hamster will, hence, forage for food when the relevant visual input to *S* is "positive", even if its low blood glucose level does not result in the secretion of orexin above the required threshold. Similarly, the hamster may also forage for food when its blood glucose level drops (and orexin is secreted above the relevant threshold as a "positive" input to *S*) without receiving the relevant visual input.[11] If so, *S* in the golden hamster may likely compute disjunction. Thus, the specific rodent-environment interaction plays a key role in fixing S' computational identity: the computational inputs and outputs need not be realised exclusively within the computing system itself. (See more in Fresco (2021, sec. 4.5).)

6.5.3 *A Midway Between the Wide Short-Arm and Mechanistic Strategies*

At this point, the astute reader may reasonably object and claim that this toy example can be explained by a short-arm mechanistic strategy; this, however, is not necessarily so. The main reason for that is that the visual inputs to *S* in both the hopping mouse and the golden hamster in response to seeds should, on average, be produced by seeds, and not by *any* light reflected from seed-like objects. This is part of the standard consumer teleosemantic story (Millikan, 1993). That is, the subsystem *S* has the *adapted proper function* of searching for food in that environment.

Given the particular environmental conditions (internal: glucose blood level, and external: availability of seeds), *S* has the adapted proper function of producing a seek-food signal. It also has the *derived* proper function of enabling the mouse and the hamster to survive in their environment by reaching out to the observed seed or

[10] Despite possible minute differences between *S* in the golden hamster and in the hopping mouse, what matters here are the input-output relations and the connectivity between *S* and the relevant upstream/downstream subsystems. Thus, even if to qualify as a "positive" input to *S*, the orexin threshold is slightly higher, say, in the mouse (as compared to the hamster), this difference is not functionally important. Such differences may manifest even between different mice of the *same species*. For similar reasons, we do not doubt that the hypothalamus as a neuroendocrine organ exists in both the hamster and the mouse despite any physical differences between them.

[11] It is probably for that reason, that vets often recommend to make home-grown hamsters work hard for their meals and hide food pellets or seeds inside paper bags or cardboard tubes.

even by seeking yet-unseen seeds (in the hamster).[12] The kinds of objects that cause *S* to compute are usually a part of the environment of the organism, and hence, the inputs may be long-arm. Likewise, the kinds of effects of *S′* computation may affect the environment, and hence, *S′* outputs may be long-arm.

A consequence of adopting this long-arm, functional individuative strategy, however, need not amount to a full-blown externalist strategy.[13] Why? Whether a physical state *p* having some proper function suffices for *p* to also represent some feature, event or object in the environment depends on the relevant theory of representation that one adopts (Fresco, 2021, sec. 4.4). Indeed, if, one is very liberal about what it takes for physical states of an organism to represent (Millikan, 1989),[14] then the proposed long-arm, functional individuative strategy entails that system *S* in the rodents does not only have the proper function of yielding seek-food commands, but it also represents the presence of food (or some such). The result, then, is a teleo*semantic* individuative strategy of computation.

On the other hand, if one adopts a more restrictive view of representation, then the long-arm individuative strategy need not amount to a full-blown externalist strategy. One such restrictive view of representation is Lloyd's (1989). According to his view, it is insufficient for *p* to yield some behavioural output (e.g., a seek-food command in our rodents) in order for *p* to qualify as a representational vehicle. Another example is Sterelny's account of representation (1995). On this account, the physical states of *S* need not qualify as representations either, because simple control systems need not amount to being representational of the very events or features that they control. One last example is Schulte's account of representation (2015), according to which constancy mechanisms are needed in addition to the function of a given system to track some environmental feature. The existence of such constancy mechanisms in the rodents is not posited by the long-arm individuative strategy. Extending all the way into the environment does bring the long-arm strategy closer to the semantic one, but it certainly does not quite go all the way there.

The take-home message is that *S* has two different proper functions depending on how *S′* computation is affected by and contributes to the organism-environment interaction. In the hopping mouse, *S* has the function of triggering a seek-food motor command *iff* both inputs are "positive" (i.e., to compute conjunction on the inputs).

[12] A seed-like object with similar surface properties of a seed may be further discriminated by the rodent's main olfactory system, which influences its foraging behaviour and food preferences.

[13] Dewhurst indeed raises this objection against Piccinini's short-arm mechanistic strategy. He claims that it is not clear how Piccinini's strategy avoids the risk of being equated with a semantic theory of computation. For "once we have teleological functions we are not far from having a full-blown teleosemantic theory of representation" (Dewhurst, 2016, p. 796). Coelho Mollo's individuative strategy—discussed in Sect. 6.2.2—similarly appeals to teleological function, but denies even narrow content, such as logical properties.

[14] For Millikan, a representation simply requires that the organism (or a consumer subsystem) can fulfil its task normally when the producer (such as *S* in the case of our rodents) goes into a state that correlates with a given environmental condition (e.g., the existence of seeds in the proximal environment).

In the golden hamster, however, *S* has the function of triggering a seek-food motor command if at least one input is "positive" (i.e., to compute disjunction on the inputs). These evolutionary functions should have been performed often enough in the evolutionary past of the respective species to have been selected. The proposed long-arm individuative strategy can tell the two apart.

At the very least, the present analysis hopefully renders the long-arm strategy plausible and a possible competitor to the wide, short-arm mechanistic strategy, on the one hand, and the full-blown externalist strategy, on the other hand. If, indeed, there are cases of computational indeterminacy that a wide, short-arm mechanistic strategy cannot adequately settle, but the long-arm functional strategy can, then the latter corresponds to yet another level in Lee's pluralistic view discussed above. Nevertheless, the long-arm strategy has to be further regimented and should include an explanation of how it applies to artificial computing systems, too.

Before concluding this chapter, an objection based on a pluralistic approach to computational explanation is briefly discussed in the next section.

6.6 A Pluralistic Approach to Computational Explanation

A possible objection might be that looking for a single, one-size-fits-all individuative strategy of computation is misguided. That is because explanatory answers to explanatory questions are generally context-dependent.[15] This objection follows the general recipe for mechanistic explanation: scientists first fix the (computational) phenomenon, and then discover its underlying mechanism(s) in a to-and-fro manner in describing the phenomenon and its mechanism (Craver, 2007). Accordingly, in looking for a computational mechanism, one should apply the suitable individuative strategy *relative* to the explanatory context. Not all computational explanations follow the same pattern of individuation, since phenomena are fixed in different manners. The suitable individuative strategy is determined in an analogous manner to bottoming out in mechanistic explanation: *where* the explanation bottoms out depends on the relevant scientific context of enquiry. If that is the case, then the long-arm functional strategy has no epistemic privilege over the others.

The response to this objection is twofold. First, it should be noted that accepting the long-arm functional strategy does not entail the denial of Coelho Mollo's individuative strategy, for example, or Piccinini's wide, short-arm functional strategy. Rather, the claim is that insofar as there exist cases of computational indeterminacy that these strategies cannot settle, the long-arm functional strategy can be invoked to determine the system's computational identity. The proposed long-arm functional strategy is not ontologically privileged over Piccinini's wide, short-arm functional strategy, for example. Thus, this approach is compatible with the pluralistic view discussed in Sect. 6.4.

[15] This interesting objection was suggested by Marcin Miłkowski.

Second, it may certainly be true that in some computational models of cognitive phenomena, scientists may avoid ambiguities in their computational hypotheses "by including semantic constraints in the specification of the explanandum phenomenon" (Miłkowski, 2017, p. 15). These are cases in which computation is performed over representations, and they are very important in cognitive science. But such cases certainly do not entail that all computation should be semantically individuated. The above toy example shows that whilst a wide, short-arm functional strategy may not be enough for its computational individuation, the long-arm functional strategy is. If that is so, then we need not appeal to further semantic properties for computational individuation.

6.7 Conclusion

This chapter examines two main approaches to computational individuation: mechanistic and semantic. Amongst proponents of the mechanistic view of computation, some advocate a very narrow individuative strategy that relies only on physical properties, whereas others appeal to contextual factors. Nevertheless, even computational mechanists—who accept that contextual factors may play a role in determining the computational identity of a physical system—insist that the functional individuation is rather narrow (i.e., it does not exceed the external boundaries of the physical system). The computational semanticist agrees that narrow contextual factors may settle some cases of indeterminacy, but argues that some interesting cases remain unanswered. She, therefore, claims that, at least in those cases, one must appeal to semantic content for computational individuation. The long-arm, functional individuative strategy may address such open cases of computational indeterminacy without adopting full-blown external content. As such, it opens up the possibility of midway between these two opposing positions.

References

Anderson, B. L. (2015). Can computational goals inform theories of vision? *Topics in Cognitive Science, 7*(2), 274–286. https://doi.org/10.1111/tops.12136

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology, 22*(5), 543–564. https://doi.org/10.1080/09515080903238948

Bickle, J. (2015). Marr and Reductionism. *Topics in Cognitive Science, 7*(2), 299–311. https://doi.org/10.1111/tops.12134

Bishop, J. M. (2009). A cognitive computation fallacy? Cognition, computations and panpsychism. *Cognitive Computation, 1*(3), 221–233. https://doi.org/10.1007/s12559-009-9019-6

Blokpoel, M. (2018). Sculpting computational-level models. *Topics in Cognitive Science, 10*(3), 641–648. https://doi.org/10.1111/tops.12282

Brooks, D. S., & Eronen, M. I. (2018). The significance of levels of organization for scientific research: A heuristic approach 1. *Studies in History and Philosophy of Science. Part C: Studies*

in History and Philosophy of Biological and Biomedical Sciences, 68–69, 34–41. https://doi.org/10.1016/j.shpsc.2018.04.003

Buckley, C. A., Schneider, J. E., & Cundall, D. (2007). Kinematic analysis of an appetitive food-handling behavior: The functional morphology of Syrian hamster cheek pouches. *Journal of Experimental Biology, 210*(17), 3096–3106. https://doi.org/10.1242/jeb.003210

Buller, D. J. (1993). Confirmation and the computational paradigm (or: Why do you think they call it artificial intelligence?). *Minds and Machines, 3*(2), 155–181. https://doi.org/10.1007/BF00975530

Coelho Mollo, D. (2017). Functional individuation, mechanistic implementation: The proper way of seeing the mechanistic view of concrete computation. *Synthese*. https://doi.org/10.1007/s11229-017-1380-5

Coelho Mollo, D. (2019). Are there teleological functions to compute? *Philosophy of Science, 86*(3), 431–452. https://doi.org/10.1086/703554

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). MIT Press.

Dewhurst, J. (2016). Physical computation: A mechanistic account. *Philosophical Psychology, 29*(5), 795–797. https://doi.org/10.1080/09515089.2016.1150450

Dewhurst, J. (2018). Individuation without representation. *The British Journal for the Philosophy of Science, 69*(1), 103–116. https://doi.org/10.1093/bjps/axw018

Fresco, N. (2021). Long-Arm Functional Individuation of Computation. *Synthese*. https://doi.org/10.1007/s11229-021-03407-x

Fresco, N., & Miłkowski, M. (2019). Mechanistic computational individuation without biting the bullet. *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axz005

Fresco, N., Copeland, B. J., & Wolf, M. J. (2021). The indeterminacy of computation. *Synthese*. https://doi.org/10.1007/s11229-021-03352-9

Gabbiani, F., Krapp, H. G., Koch, C., & Laurent, G. (2002). Multiplicative computation in a visual neuron sensitive to looming. *Nature, 420*(6913), 320–324. https://doi.org/10.1038/nature01190

Haimovici, S. (2013). A problem for the mechanistic account of computation. *Journal of Cognitive Science, 14*(2), 151–181.

Harbecke, J., & Shagrir, O. (2019). The role of the environment in computational explanations. *European Journal for Philosophy of Science, 9*(3), 37. https://doi.org/10.1007/s13194-019-0263-7

Hardcastle, V. G., & Hardcastle, K. (2015). Marr's levels revisited: Understanding how brains break. *Topics in Cognitive Science, 7*(2), 259–273. https://doi.org/10.1111/tops.12130

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science, 2*(1), 119–135. https://doi.org/10.1007/s13194-011-0038-2

Jones, P. W., & Gabbiani, F. (2012). Logarithmic compression of sensory signals within the dendritic tree of a collision-sensitive neuron. *Journal of Neuroscience, 32*(14), 4923–4934. https://doi.org/10.1523/JNEUROSCI.5777-11.2012

Lee, J. (2021). Mechanisms, wide functions, and content: Towards a computational pluralism. *The British Journal for the Philosophy of Science, 72*(1), 221–244. https://doi.org/10.1093/bjps/axy061

Lloyd, D. E. (1989). *Simple minds*. The MIT Press.

Love, B. C. (2015). The algorithmic level is the bridge between computation and brain. *Topics in Cognitive Science, 7*(2), 230–242. https://doi.org/10.1111/tops.12131

Miłkowski, M. (2016). Computation and multiple realizability. In V. C. Müller (Ed.), *Fundamental issues of artificial intelligence* (pp. 29–41). Springer. https://doi.org/10.1007/978-3-319-26485-1_3

Miłkowski, M. (2017). The false dichotomy between causal realization and semantic computation. *Hybris, 38*, 1–21.

Miłkowski, M., Clowes, R., Rucińska, Z., Przegalińska, A., Zawidzki, T., Krueger, J., Gies, A., McGann, M., Afeltowicz, Ł., Wachowski, W., Stjernberg, F., Loughlin, V., & Hohol, M. (2018). From wide cognition to mechanisms: A silent revolution. *Frontiers in Psychology, 9*, 2393. https://doi.org/10.3389/fpsyg.2018.02393

Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy, 86*(6), 281–297. https://doi.org/10.2307/2027123

Millikan, R. G. (1993). *White queen psychology and other essays for Alice*. MIT Press.

Papayannopoulos, P., Fresco, N., & Shagrir, O. (2022). On Two Different Kinds of Computational Indeterminacy. *The Monist, 105*(2), 229–246. https://doi.org/10.1093/monist/onab033

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford University Press.

Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.

Schulte, P. (2015). Perceptual representations: A teleosemantic answer to the breadth-of-application problem. *Biology and Philosophy, 30*(1), 119–136. https://doi.org/10.1007/s10539-013-9390-2

Shagrir, O. (2001). Content, computation and externalism. *Mind, 110*(438), 369–400. https://doi.org/10.1093/mind/110.438.369

Shagrir, O. (2020). In defense of the semantic view of computation. *Synthese, 197*(9), 4083–4108. https://doi.org/10.1007/s11229-018-01921-z

Sprevak, M. (2010). Computation, individuation, and the received view on representation. *Studies in History and Philosophy of Science Part A, 41*(3), 260–270. https://doi.org/10.1016/j.shpsa.2010.07.008

Sterelny, K. (1995). Basic minds. *Philosophical Perspectives, 9*, 251. https://doi.org/10.2307/2214221

Stinson, C. (2016). Mechanisms in psychology: Ripping nature at its seams. *Synthese, 193*(5), 1585–1614. https://doi.org/10.1007/s11229-015-0871-5

Wells, A. J. (1998). Turing's analysis of computation and theories of cognitive architecture. *Cognitive Science, 22*(3), 269–294. https://doi.org/10.1207/s15516709cog2203_1

Chapter 7
Levelling the Universe

John Heil

Abstract Reductionist, 'bottom-up', programmes in philosophy and in the sciences are often depicted as attempts to explain higher-level phenomena in lower-level terms. In practice this would mean deriving higher-level predictions and explanations from laws governing phenomena at lower-level (perhaps with the help of 'bridge principles'). Reductionism has not lived up to expectations. Scattered successes have been overshadowed by widespread failures. In response, many researchers have embraced a hierarchical conception of the universe. This chapter reflects on the history of considerations thought to support this hierarchical conception and concludes that the considerations have negligible metaphysical weight.

> A philosopher is commonly thought of as a reasoner, but I would rather conceive him as a person careful in his assumptions. (Strong, 1923: x–xi)

7.1 Prelude

Philosophical platitudes and assumptions have histories. Many gain currency in the face of difficulties thrown up by a perceived discordance between what the sciences tell us about the cosmos and the cosmos as we experience it. Scientific theories rest on observations, but what we encounter in the course of observing is apparently at odds with the picture painted by physics. Recent attempts to reconcile the appearances with reality as physics characterises it began with reductionism and its evil twin, instrumentalism, and eventually came to settle on a conception

J. Heil (✉)
Washington University in St Louis, St. Louis, MO, USA

Durham University, Durham, UK

Monash University, Melbourne, VIC, Australia
e-mail: jh@wustl.edu

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_7

of the universe as hierarchical (see, for instance, Oppenheim & Putnam, 1958). Physics affords a systematic account of the fundamental, lowest-level phenomena, the special sciences provide accounts of phenomena occupying successively higher-levels.

In what follows I offer an opinionated reconstruction of the history of the reasoning that led us to this point. (See Heil, 2021 for a more cautious and detailed treatment of topics to be addressed here.) My aim is not principally historical I am, however, convinced that understanding how we came to be as we are can lead to a reassessment of the state of play. The professionalism of philosophy with its attendant pressure to publish encourages us to take on board positions, not because those positions recommend themselves to us, but because they are widely taken for granted. If we accept what those who call the shots accept, we can attend to the details without having to reinvent the wheel.

Once you step back and reflect on factors that shaped the way we think about the cosmos and our place in it, however, you can see that our route to the present, far from being a well-maintained highway, is an inauspicious byway riddled with potholes, detours, and cul-de-sacs. Philosophy advances in fits and starts, often doubling back on itself. Our recent history is one of philosophers talking past one another and advancing arguments that rest on assumptions that pass muster only because they have become invisible by virtue of being widely shared.

Enough drama! Time to shut up and deal.

7.2 Parts and Wholes

Today, many philosophers and scientists think of the cosmos as comprising a hierarchy of 'levels', the levels roughly corresponding to the domains of the various sciences, and, ultimately everyday experience. Sometimes levels are spelled out mereologically. At the fundamental level you have quarks and leptons, which make up atoms, which make up molecular structures, which make up cells, which make up organisms, and so on for species, habitats, ecosystems, communities, societies, civilizations, galaxies, the cosmos as a whole. Higher levels encompass wholes, the parts of which, and parts of those parts, make up successively lower levels.

Interesting wholes cannot be reduced to their parts. Laws governing wholes and lawlike generalisations true of wholes are not derivable from laws and generalisations applicable to the parts. In many cases capacities and powers of the wholes outstrip those of the parts. A whole can do things that would be impossible to explain on any but an ad hoc basis by reverting to descriptions of its parts. A whole appears to be a fully fledged *it* fitted out with properties not found in its parts. And, although this is more controversial, wholes might be implicated in the behaviour of the parts: the parts behave as they do, not only because they are governed by their own laws, not only because of their individual natures, but because they belong to particular wholes.

This last point is sometimes spelled out in terms of downward, whole-to-part causation. Certain kinds of whole can exert 'configurational forces' on parts that make them up. Roger Sperry puts it this way for 'subjective mental phenomena':

> Subjective mental phenomena are conceived to influence and govern the flow of nerve impulse traffic by virtue of their encompassing emergent properties. Individual nerve impulses and other excitatory components of a cerebral activity pattern are simply carried along or shunted this way or that by the prevailing overall dynamics of the whole active process (in principle—just as drops of water are carried along by a local eddy in a stream or the way the molecules or atoms of a wheel are carried along when it rolls down the hill, regardless of whether the individual molecules and atoms happen to like it or not). (Sperry, 1969: 534)

Talk of whole-to-part causation is associated with talk of emergence. Emergent phenomena are distinguished by the fact that they amount to something new on the scene, something requiring the introduction of new laws and new powers.

More is required for capital-*E* Emergence, however. A ball assembled from Lego bricks has the power to roll owing to its spherical shape, although none of the bricks that make it up have this power. But it is easy to see how you get from non-spherical bricks to a spherical object with the power to roll. The ball's sphericality is a merely 'resultant' property of the ball.

An emergent property, in contrast, would amount to more than a whole's possessing a property equipping it with powers not possessed by its parts. An emergent property equips a whole with a power to affect the behaviour of its parts, typically in ways that figure in the whole's ongoing integrity in the face of environmental adversities. This is the feature of emergent properties that separates them from merely resultant properties.

7.3 Reduction and Identity

Before saying more about emergence, a second route to the hierarchical picture deserves mention. This route is perhaps taken more often by philosophers than by scientists, although it, too, begins with a strong antireductionist commitment. I have in mind the nonreductive physicalist programme kicked off in the 1960s by philosophers grappling with the mind–body problem.

The late 1950s saw the rise of materialism in the form of the mind–brain 'identity theory' advanced by U. T. Place and J. J. C. Smart, both (together with C. B. Martin) then at the University of Adelaide (see Place, 1956; Smart, 1959). According to the identity theory, mental states and processes are, as a matter of empirical fact, brain states and processes. Although the mental is not conceptually reducible to the physical, mental properties are nevertheless identified with, and in that regard reducible to, physical properties.

The identity theory begins with the observation that mental phenomena—or, at any rate, subjects' reports of mental phenomena—are correlated with goings-on in the brain. Dualism, which requires positing two distinct kinds of property,

is saddled with the problem of explaining the correlations. Do minds and bodies interact? If so, how would such interaction work? How could a purely physical system affect a nonphysical system? How could something physical get a grip on something nonphysical? And how could anything nonphysical intervene in physical causal processes? Or do minds and brains miraculously operate in parallel? When you bark your shin and subsequently experience a painful sensation, is this because your mind and body are, like two clocks keeping time in tandem, marching in step?

Recognising the prima facie implausibility of parallelism and the difficulty of explaining causal interactions among physical and nonphysical phenomena, some theorists embraced epiphenomenalism: mental phenomena are by-products of physical processes, much as the heat produced by the operation of a machine might be a by-product of the machine's operation. The heat is produced by the machine, but plays no role in the machine's operation.

One advantage of epiphenomenalism is that, if true, it would relieve the physical sciences of having to reckon with mental phenomena in physical explanations. If the mental had no affect on the physical, it could be safely ignored. In a similar vein, at the start of the scientific revolution in the seventeenth century, scientists—natural philosophers—dispensed with secondary qualities, colours, sounds, tastes, smells, and the like by relegating them to the minds of observers. By placing these outside the physical fray, natural philosophers could disregard them with a clear conscience.

Once you start worrying about the relations states of mind bear to their physical correlates, however, you again come face-to-face with the mind-body problem. If you are interested exclusively in physical phenomena, you might get away with locating secondary qualities in the minds of observers. No such move is available when you turn your attention to minds and their contents.

Materialism, in the guise of the identity theory, offered a no-nonsense solution to the difficulty. Mental properties *are* physical properties. The correlation between the mental and the physical is specious, in the way the correlation between orbits of the Hesperus and Phosphorus is specious. 'Hesperus' and 'Phosphorus' are two names for one and the same heavenly body. There is just the one orbit.

7.4 Functionalism

Advocates of mental-physical dualism were predictably unhappy with the identity theory, chiefly on the grounds that it was implausible to think that qualitatively rich conscious phenomena are nothing more than drab physical occurrences in the brain. At the time, however, the most serious threat to the mind–brain identity theory was not dualism, but functionalism.

Hilary Putnam and Jerry Fodor, two important figures in the rise of functionalism, argued that identifying mental states with states of the brain was a kind of category mistake analogous to the mistake of identifying the program running on a computing machine with physical states of the machine (see Putnam, 1967; Fodor, 1981). The same program could run on machines that were altogether different physically.

Minds are programs implemented in neural hardware. Functionalism is consistent with, but does not entail materialism.

Consider what it is to be in pain. To be in pain is to be in a state with a particular causal profile. A pain state is a state caused by tissue damage that itself causes aversive reactions. (A functional analysis of pain would in fact be much more complicated, but this will do by way of illustration.) Contrary to the identity theory, the property of being in pain, is not the property of being in a particular brain state. Different species of creature could experience pain despite having very different physical constitutions. The property of being in pain—the pain property—must be a 'second-order' property. The pain property is the property of having a first-order property with the right causal profile.

A second-order property in this context is not what you might think. A second-order property is not a property of a property. It is the property of having a particular first-order property. For that reason, I prefer to refer to the properties in question as higher-*level* properties.

Functionalists take the pain property, and mental properties generally, to be higher-level properties possessed by a creature by virtue of that creature's possession of the right sort of lower-level physical properties. Higher-level properties and states were said to be 'realised by' lower-level properties and states. Higher-level properties were presumed to be distinct from—not reducible to—but nevertheless dependent on their lower-level realisers. The only way to get a higher-level property on the scene would be to bring it about that an appropriate lower-level realiser is on the scene.

This conception of property levels was extended to cover scientific domains that had nothing to do with minds. Biological properties would seem to be higher-level properties with cellular realisers. These, in turn, had molecular realisers, and so on until you reached the level of quarks and leptons. The hierarchy of levels extends 'upwards' as well, the result being an edifice with its upper stories in the clouds, supported by a stolid physical foundation.

This picture is not unlike what you have in the case of wholes and parts. The approaches are united by a shared commitment to antireductionism and to the idea that higher-level entities are entities in their own right. The reason the various sciences do not reduce to sciences at lower levels is that higher-level sciences are concerned with higher-level domains of phenomena that answer to their own gods.

7.5 Appearance and Reality

Before looking more closely at hierarchical pictures of the cosmos, I propose to step back and try to put all this talk about levels and hierarchies into perspective. I have mentioned two ways of fitting the various sciences together. First, there are the reductionists who see the sciences as ultimately unified through reduction. We compartmentalise the sciences, not because they concern distinct realms of phenomena, but solely for convenience. The complexity of larger systems forces

us to operate with approximations and less fine-grained categories that serve well enough, but are in principle, dispensable. When you get down to business, it is all just physics.

The failure of reductionist programmes bred the hierarchical, levels picture. Reduction does not work because the several sciences are concerned with several realities. These are organised hierarchically, with those at higher levels being dependent on those at lower levels.

A third approach, instrumentalism, deserves mention. Confronted with distinct accounts of the springs and levers underlying the observed phenomena, instrumentalists argued that, despite appearances, the sciences are not in the business of uncovering what lies behind observation. Rather, the sciences provide systematic ways of negotiating what is observed. Owing in part to human cognitive limitations, these often take on a concrete form. If we operate as though the universe comprises quarks, leptons, and gravitational fields, for instance, we can sharpen our predictions. Scientific theories are black boxes. You feed in one set of observations, and the box converts these to predictions that can be checked against new observations. To the extent that these pan out, the theory is confirmed. When a theory's predictions are not confirmed, the theory is adjusted accordingly.

Instrumentalism

You could see all three of these—reductionism, instrumentalism, and the hierarchical picture—as offering answers to the question, how are the appearances related to reality? In speaking of 'appearances' here, I mean to be singling out, not merely the way things strike to us unreflectively as we go about our everyday pursuits, but also the way things appear to scientists in their laboratories working with mass spectrometers and cloud chambers, or in the field studying insects or iron age agriculture. The upshot is three approaches to the question, how are the appearances related to reality?

1. The appearances are *mere* appearances; reality is what issues from the sciences, ultimately physics (reductionism).
2. The appearances are what is real; physics and the other hard sciences are in the business of devising constructs the sole purpose of which is to facilitate our give and take with the appearances (instrumentalism).
3. The appearances comprise a hierarchy of domains or levels in which higher-levels depend on lower-levels, ultimately bottoming out in a ground-floor level, the domain of physics (levels of reality).

I believe that all three of these attempts to resolve the tension between the cosmos as it appears to us and the cosmos as characterised by physics are unsatisfactory. As I hope the foregoing makes clear, my discussion is intended to be suggestive,

not exhaustive. The issues that concern me here have been addressed in considerable detail in many places by many philosophers. Anyone acquainted with that literature, including proponents of levels, would likely grant that the case for levels remains inconclusive at best. My contention is that, accepting a metaphysics of levels requires compromises, evasions, and circumlocutions that exude an air of desperation.

7.6 Emergence and Downward Causation

Emergent phenomena would qualify as higher-level, but what exactly is an emergent phenomenon? Some accounts of emergence dwell on the difficulty or impossibility of deriving truths concerning wholes from truths about their parts. By 'truths', I mean not only statements or descriptions, but also laws and lawlike generalisations of the sort familiar in the special sciences.

Part of the problem stems from the fact that concepts and categories at higher-levels crosscut those at lower levels. As a result, what count as entities at higher levels appear ad hoc or gerrymandered from the perspective of lower levels. Entities at higher levels, their properties, and any laws in which they might figure fail to align with their lower-level counterparts.

Emergence, so construed, is sometimes called weak emergence owing to its epistemological character. Weak emergence amounts to the claim that higher-level truths cannot be known on the basis of lower-level truths. I suspect that this is what many scientists have in mind when they speak of emergence. Evidence for weak emergence appears to be overwhelming, so I shall not challenge it here. What I shall challenge is the move from weak emergence to the levels picture.

To move from emergence to levels you need *strong* emergence. I admit that I am unclear what exactly strong emergence amounts to. In § 2, I suggested that there appears to be a connection between this strain of emergence and downward, whole-to-part causation. A strongly emergent whole would be one capable of exercising authority over its parts. You have parts—quarks and leptons, or molecules, or cells, or organisms—making up wholes that, once on the scene, themselves influence the behaviour of the parts.

Whether there are any wholes capable of exerting forces on their parts, is an empirical question. Note, however, that appealing to epistemological arguments of the kind that support weak emergence, would not be to the point. Something more is required, and I am sceptical that we have anything approaching uncontested empirical grounds for that something more.

Those of us not already committed to strong emergence, find it hard to understand what it would be for a whole made up of parts to influence those parts causally. Are emergent wholes *causa sui*, entities made up of parts that can causally influence themselves? The impression that some organised wholes causally influence the behaviour of their parts might simply be one consequence of selling the parts short. There is no question that parts of familiar organised wholes can influence

one another, often in ways that result in the parts undergoing changes. There is no question that parts in a particular arrangement behave differently than they would in a different arrangement. But it is a leap from this to the idea that the arrangement is responsible for the arranging.

Think of Sperry's 'local eddy in a stream' (introduced in § 2). Sperry deploys the example as an instance of top-down causation in which 'drops of water are carried along' by the eddy 'regardless of whether the individual molecules and atoms happen to like it or not'. Jaegwon Kim speaks for many when he observes that 'an eddy is there because the individual water molecules constituting it are swirling around in a circular motion; in fact, an eddy is nothing but these water molecules engaged in this pattern of motion' (Kim, 2000: 313).

There is more to be said here, but this is not the place to say it. Please understand that I am not claiming that emergence—strong emergence—is impossible or that it can be ruled out a priori. My aim is only to indicate that emergence and downward, whole-to-part causation are more puzzling than proponents of emergence might want you to think. Strong emergence is motivated, at least in part, by under-describing interactions among the parts, and by a tendency to move from the undoubted fact that parts behave differently in different arrangements, to the idea that the arrangements themselves have a causal role in the arranging.

7.7 Quantum Holism

A word is in order concerning the kind of holism you find in quantum physics. Particles in entangled states do not interact in the purely mechanical fashion I have suggested is responsible for the characteristics and behaviour of familiar complex wholes. Particle trajectories reflect the trajectories of fellow particles in ways that cannot be explained causally. The behaviour of individual particles in entangled states appears explicable only by reference to the whole collection of particles. Might this be evidence for emergent entities and downward causal influence?

A better way of thinking about such cases conceives of the particles, not as mereological parts of wholes, but as abstract particulars, abstract, not in the sense of being non-spatiotemporal, but in the sense of being particular modifications of a grander something—a field, perhaps, or spacetime, or the cosmos itself. Modifications, what were traditionally called modes, belong to the category of property, not substance.

An object's parts, but not its properties, interact causally. In the case of quantum entanglement, what we think of as individual particles moving about and interacting with one another are analogous to a shiver's running down your spine. The analogy is limited, however. The shiver is a causal product of occurrences inside your body, but the apparent motions and interactions among particles are not caused by the whole, they are expressions of the whole's dynamic nature.

I discuss this cosmology in more detail elsewhere (Heil, 2021). I mention it here simply to forestall concerns that I myself am guilty of substantive but

unacknowledged assumption: a corpuscular cosmos. I have elected to carry on the discussion against a corpuscular background, not because I accept it, I do not, but because it simplifies the presentation without begging the question against proponents of levels.

7.8 Abandoning Reduction

In § 2, I observed that a second route to a hierarchy of levels originates with arguments against reduction in the 1960s and 1970s, in the first instance, the reduction of mental states to brain states, and then the reduction of phenomena dwelt on by the special sciences to phenomena investigated by lower-level sciences. Having myself partaken of the cup, I can appreciate the pull of the arguments. Philosophers, like everyone else, are subject to a variety of influences that work in the background. Indeed, their efficacy largely depends largely on their remaining in the background.

Once a position is taken up by a critical mass of philosophers, it becomes difficult not to accept the position as a given, and proceed from there. In my own case, I can recall being puzzled by arguments for multiple realisability, but I was sure that my puzzlement arose from a failure on my part to appreciate what others, far more capable than I, did appreciate. Under the circumstances I was content to work as an underlabourer, reassured by the willingness of journals to publish what I was turning out.

Now, in retrospect, it is easier to see that the most influential arguments against reduction were broadly semantic in character. Take psychology and its relation to neuroscience. Psychological taxonomies crosscut neurological taxonomies. You cannot derive psychological truths from truths of neurophysiology. Psychological categories do not map smoothly onto biological categories, much less onto those of physics. And, as it is for psychology, so it is for all the other special sciences.

These taxonomic arguments were gripping. Problems arose, however, when those advancing the arguments proceeded to give them a metaphysical cast. Philosophers more at home in the philosophy of language than in metaphysics drew inspiration from the work of Donald Davidson (see Davidson, 1967, 1970, 1973). Davidson, for reasons having to do with his work on truth and meaning, accepted that ascriptions of propositional attitudes—beliefs, desires, and intentions, for instance—could not be translated or paraphrased in a purely physical vocabulary. Mental terms owed their significance to their membership in a cohort of mental terms. These played by their own rules, and these rules had no echo in the physical realm.

Crucially, for Davidson, the mental domain does not float free of the physical domain. It is true of Donald that he was born in Springfield, weighs 11.5 stone, and is 175 cm tall. It is true of the very same Donald that he believes Hesperus is Phosphorus, wants it to stop raining, and intends to play a Chopin étude on the piano after supper. We have these two autonomous domains of terms, the mental and the physical, that nevertheless hold true of the selfsame Donald.

More importantly, perhaps, is the apparent fact that satisfaction conditions for mental predicates essentially include causal relations to both states of mind and states outside the mind. Your belief that it is raining counts as a belief about rain (and not about what you had for breakfast) in part because its causal history includes rain, not your breakfast. Your intention to walk to the closet is the intention it is (and not some other) because it is embedded in a causal network that includes particular beliefs and desires. How might this work? Davidson appealed to supervenience. The mental supervenes on the physical: no mental difference without a physical difference.

I shall return to Davidson in due course. First, however, it behoves me to carry on with my version of the history of how we came to be where we are.

7.9 Horrible Histories

As mid-twentieth century philosophers began to move on from attempts to analyse mental terms in a physical vocabulary, metaphysics was regaining respectability. The English-speaking philosophical community was largely ill-equipped to make the transition, however. Old habits die hard. We philosophers had become comfortable with a ratbaggish linguisticized brand of metaphysics, one that allowed metaphysical categories to be read off linguistic categories. The practice is on the whole harmless so long as the metaphysics is not taken ontologically seriously. Problems arise, however, when linguistic shadows are mistaken for the real McCoy. Discussions of properties in the 1960s and 1970s are a case in point.

For years we philosophers took for granted that what it was for an object to possess a property—for a cricket ball to possess the property of redness, for instance—was for the predicate 'is red' to be true of the ball. All there is to an object's possessing a property is for a predicate to be true of the object. Predicates do not *correspond* to properties. What makes it true that the cricket ball is red is not the ball's possessing a property, redness, but simply the cricket ball itself.

When philosophers started taking properties seriously, the practice continued, the only difference being that now predicates true of objects were taken to be true because they corresponded to objects' properties. Distinct predicates true of an object must designate distinct properties. (Predicates are distinct if neither is eliminable in favour of the other, neither can be analysed or paraphrased in terms of the other.)

Now, return to Davidson's contention that the mental supervenes on the physical and allow that mental predicates are not analysable in a physical vocabulary. That being the case, mental and physical predicates must correspond to distinct families of property. To say that the mental supervenes on the physical is to say that mental properties, while patently distinct from physical properties, are nevertheless dependent on physical properties, supervenience being a label for the dependence relation.

This line of reasoning spawned an industry devoted to spelling out the metaphysical character of the supervenience relation. Taxonomies of species of supervenience were advanced, all in the service of giving supervenience a metaphysical backbone. Along the way, it became clear that supervenience was by no means restricted to mental and physical properties, but had wide application. Psychological properties might be thought to supervene on biological properties, biological properties to supervene on chemical properties, and chemical properties to supervene on properties discovered by physics. The upshot is a hierarchy of properties (and presumably property bearers) generated by the supervenience relation.

In practice, the situation was more complicated, but the eventual result was a hierarchical conception of the universe according to which properties that belong to the domains of higher-level sciences were taken to supervene on properties belonging to the domains of lower-level sciences until you reached the domain of physics. Many came to accept this picture as mandated by the sciences.

By this time supervenience had been recruited into the service of multiple realisability. Mental properties have physical realisers, but supervenience allows for the possibility that one mental property—the pain property, for instance—could have many different physical realisers. This fits nicely with functionalism and the idea that mental properties depend on, but are not reducible to physical properties.

A loose end remained. Although supervenience was widely invoked, there was little or no agreement on the metaphysical details. Supervenience was presumed to be a kind of dependence relation, but the nature of the dependence remained obscure. Kim, who had devoted years to attempts to get clear on species of supervenience, came to the conclusion in the 1990s that supervenience, as commonly characterised, boiled down to covariation among property families. And, as he observed, more than covariation, even necessary correlation, is required to fill out the dependence relation gestured at by supervenience (Kim, 1990).

Still, we remained in the grip of the hierarchical, levels picture. If philosophers find the metaphysics of supervenience (or realisation) mysterious, so much the worse for the philosophers. Our best science tells us that mental properties supervene on physical properties, indeed, legions of higher-level properties supervene on properties belonging to lower levels. That is a given. Who are philosophers to question the sciences? Maybe supervenience is at bottom a *sui generis* metaphysical relation not further explicable. If that is how it is, so be it.

7.10 Causal Relevance

We thus arrived at a hierarchical metaphysical picture riddled with caveats. Chief among these stemmed from the difficulty of understanding how higher-level supervenient properties could have causal relevance. On the one hand, the supervenient status of higher-level properties, including mental properties, seemed unchallengeable. On the other hand, many accepted Plato's advice to take causal efficacy as the mark of reality: to be real is to be causally efficacious. In that case, if

mental properties—and higher-level properties, generally—are real they must be empowering in distinctive ways not reducible to those of their realisers.

The sciences, including psychology and all the other special sciences are rife with talk of causal interactions among higher-level phenomena and between these and lower-level phenomena. Given the dependence of higher-level properties on their lower-level supervenience base, however, it was unclear how higher-level properties could equip their possessors with causal powers over and above those stemming from the realising properties.

You can see the problem by imagining a concrete case. Suppose you look out of the window and see that it is raining. In consequence you form the intention to retrieve your umbrella from the hall closet and straightaway proceed to the closet. In this instance, your belief that it is raining causes you to form the intention to retrieve your umbrella, as a result of which you walk to closet across the room. Yes, but how could your intention, a higher-level state, mobilise your body? The physical realiser of your intention would seem to be what does the work. In that case, your intention would not itself be relevant to the production of your subsequent behaviour. The nature of mental-to-physical—and, in general, higher- to lower-level—causation remained mysterious.

Perhaps higher-level items causally interact exclusively with other higher-level items: your belief that it is raining causes your intention to look for your umbrella. Even if higher-level states cannot affect lower-level states, higher-level states might be able to affect other higher-level states: intra- not inter-level causation.

Difficulties abound. Your belief is on the scene because its physical realiser is on the scene. This is a straightforward consequence of the one's supervening on the other. But now how would your belief—a higher-level state—cause your intention, another higher-level supervenient state? Your intention is itself dependent on its own lower-level realiser. Your intention's being on the scene requires that its realiser is on the scene. Your belief's causing your intention would require your belief's bringing about your intention's lower-level realiser. A higher-level state—your belief—could bring about another higher-level state—your intention—only by bringing about its lower-level realiser.

Now, as Kim noted, the problem is that if your belief, a higher-level state, causally brings about the lower-level state that realises your intention, you have higher-level items intervening in lower-level causal sequences. Given that the levels eventually bottom out at the level of the quarks and leptons, this would mean that your belief, a higher-level state, intervenes in ground-level causal networks in a way not unlike that suggested by Sperry.

Although this kind of downward causal influence cannot be ruled out a priori, there are empirical difficulties of the kind pointed out by physicist, Carlo Rovelli:

> There is nothing about us that can escape the norms of nature. If something in us could infringe the laws of nature, we would have discovered it by now. There is nothing in us in violation of the natural behavior of things. The whole of modern science—from physics to chemistry, and from biology to neuroscience—does nothing but confirm this observation. (Rovelli, 2016: 72–73).

Is Rovelli philosophically naïve, influenced, perhaps, by outdated reductionist prejudices? The hierarchical picture is, after all, dictated by the sciences. If the picture requires downward causation, our only option is to accept it. Philosophers should stick with philosophy and not meddle in the sciences.

7.11 Anomalous Monism

So here we are with an imposing hierarchical edifice, one with the imprimatur of the several sciences and vouched for by respected and influential philosophers. Metaphysical difficulties, if there are any, fall to the philosophers; the scientists have better things to do.

Is this right? Is the hierarchical edifice a scientific given? I doubt it. Although many factors were in play in the 1980s and 1990s, I believe the levels picture was a product of an ill-considered metaphysics rooted, at least in part, in a fundamental misreading of Davidson.

In arguing that the mental supervened on the physical, Davidson was interpreted as making a metaphysical claim concerning families of property: mental properties supervened on physical properties. The upshot was an eruption of work on mental causation, aimed at answering, or at least defusing, the question, how could supervenient properties enter into causal relations? The efficacy of higher-level properties was apparently undercut by the physical realisers of those properties.

When doubts were expressed about this dialect, many stepped forward noting that what goes for mental properties, goes for properties at home in the special sciences. A problem faced by everyone, is no one's problem.

As someone profoundly influenced by Davidson, I have finally gained the confidence to step back and reconstruct what was happening. I do so with considerable humility. My reconstruction reflects my own philosophical trajectory. Even if I am philosophically unrepresentative, however, I believe that there are lessons here for anyone attracted to the hierarchical, levels picture.

The first point to note is that Davidson, a student of Quine's, was comfortable using 'property' as a stand-in for 'predicate'. When Davidson claimed that mental properties and physical properties were distinct, what he meant was that truth conditions for the application of mental predicates were orthogonal to truth conditions for applications of physical predicates. Mental predicates, and psychological descriptions generally, could not be analysed or paraphrased in a physical vocabulary. Like Putnam, Davidson had in mind the failure of the behaviourist programme in its efforts to provide analyses of mental terms in a nonmental vocabulary. In this he differed from Quine.

Davidson, then, was happy to accept talk of mental properties supervening on physical properties. But Davidson's use of 'property' reflected an ontologically recessive use by his generation of philosophers. To say that a cricket ball as the property of being red is to say no more than that the cricket ball—*holus bolus*—

satisfies the predicate ‘is red’. There need be no characteristic, no feature, no aspect of the cricket ball corresponding to the predicate.

In this context, I think it best to read Davidson, not as offering a defence of nominalism, actively denying that objects have properties in an ontologically serious sense. Rather, it simply did not occur to him that mental and physical predicates designated distinct families of ontologically robust properties. He was interested in understanding the relation between two different ways of describing the cosmos: the mental way and the physical way.

Second, Davidson accepted that causal relations are relations among events that fall under strict, exceptionless laws. There are, he argued, no such laws when it comes to mental–physical interactions, however. Causal laws are expressible only in an exclusively physical vocabulary. In failing to be physically reducible, the mental is anomalous. How, then, could there be causal relations among mental and physical events? Without such causal interactions, it would be impossible to account for the evident success of our practice of explaining actions by ascribing states of mind to agents.

Answering this question requires an appreciation of a third reason to doubt that, in invoking supervenience, Davidson was advancing a metaphysical thesis about dependence relations among families of property. Davidson inherited talk of supervenience from R. M. Hare who had argued that moral truths supervene on nonmoral, natural truths (Hare, 1952, 1984). Hare, himself, was an antirealist about normativity. His idea was that, although moral truths cannot be deduced from nonnormative, physical truths, normative ascriptions hold of agents by virtue of those agents’ nonnormative features.

Davidson appreciated that supervenience brought with it no commitment to anti-realism. Supervenience as Davidson conceived of it has two aspects.

1. Antireductionist arguments provide evidence for taxonomic incommensurabilities. Davidson accepts this conclusion for psychological categories: these are not reducible to—replaceable by—physical categories. There is no principled mapping between the application conditions of mental and physical predicates. In this regard psychology is an autonomous discipline, not replaceable by biology or neuroscience.
2. Although mental truths could not be captured in a physical vocabulary, whatever answered to a mental predicate, answered as well to some physical predicate. Whenever you have something that could be given a mental description, it could be given a physical description, although it is no part of the view that you must be in a position to provide the physical description. Truth conditions for the application of mental predicates do not map smoothly onto those for physical predicates.

Supervenience, as Davidson understood it, is a substantive doctrine supported by the fact that one and the same agent answers to both mental and physical predicates, and by its providing an explanation of mental–physical causal interactions. I addressed the first of these in § 7.8, but the second takes us to the crux of the matter. When your intention to walk to the closet to retrieve your umbrella leads you to walk, your

walking is caused, in part, by your having formed the intention. How might that work? More generally, how is mental–physical causal interaction possible?

If the mental supervenes on the physical, every mental state, every state answering to a mental predicate, is a physical state, a state answering to a physical predicate. I have insisted that, unlike many philosophers who read him, Davidson did not think of the mental and the physical as different species of property, one depending on the other. In fact, Davidson's brand of supervenience goes both ways. Anything that could be given a mental description could be described in a physical vocabulary, and vice versa: anything truly describable in a physical vocabulary could be picked out using a mental vocabulary. What you have, in essence are two distinct ways of describing the same things: mental–physical monism. Because these distinct ways are not in alignment, however, Davidson's monism is anomalous: anomalous monism.

When your forming an intention brings about your walking, this is not because the sequence is an instance of a strict law connecting intentions to bodily motions. Owing to the misalignment of mental and physical predicates, there are no such laws. If you accept supervenience, however, you except that your mental state answers to some physical description, and this, together with a physical description of your bodily motions, would constitute an instance of a strict law—whether or not you could formulate it.

Philosophers who took supervenience to be a relation among families of property dismissed Davidson's account of mental causation as hopeless. They assumed that Davidson accepted that mental properties were distinct from, yet dependent on physical properties. Davidson's suggestion that every state that answered to a mental predicate also answered to a physical predicate, was reconstrued as a claim about properties: every event with a mental property has a subvenient physical property. (I am being deliberately vague in speaking of mental and physical properties of events. Were these properties of events, or constituents of events? Different philosophers told different stories.)

Thus, even if it is true that whenever a mental event—an event with a mental property—caused a physical event, the mental event has a subvenient physical property, and it is this physical property that figures in a causal law. Yes, the mental event has a mental property, it is after all a mental event. Its causing a physical event was due, not to this property, however, but due to some subvenient physical property. Mental properties, although invariably accompanied by physical properties, were causally irrelevant, epiphenomenal.

The criticism, regarded by many as conclusive, utterly missed the point. According to Davidson, if the mental supervenes on the physical, this is not because mental properties supervene on nonmental, physical properties. Supervenience is not a relation among property families but a relation among families of predicate. 'Physical' is not to be read as 'nonmental', nor is 'mental' to be read as 'nonphysical'. One and the same state can answer to both mental and physical predicates, one and the same state can be both mental and physical.

Supervenience amounts to the idea that, although mental truths could not be expressed in a physical vocabulary, whatever answers to a mental predicate, answers

as well to a physical predicate. A state is mental or physical, 'only as described' (1970: 89, 1980: 215). To ask whether an event caused another event because of its being mental or its being physical, would be to ask whether an event caused another event because it answers to a mental predicate or because it answers to a physical predicate, a question Davidson rightly regarded as confused. No, you have just the one event answering to distinct predicates.

7.12 Dénouement

You might not be on board with Davidson's position as I have described it, but I hope at least to have convinced that, whatever its defects, it avoids the metaphysical pitfalls of a hierarchical metaphysics. More than that, I believe it comports nicely with scientific practice. The idea that irreducibility entails a hierarchical metaphysics of levels does not emanate from the sciences, but from a philosophically tendentious reconstruction of the sciences.

Philosophers like to speak of science as carving reality at the joints. Reality has many, maybe uncountably many, joints and admits many carvings. The several sciences are distinguished from one another by their distinctive ways of addressing the cosmos as we find it. There is but one cosmos the occupants of which can be understood, described, and explained in many distinct, nonequivalent ways. If there is a hierarchy, it is a hierarchy among these different modes of understanding, description, and explanation.

You can describe something's parts, and you can ignore the parts and describe the whole they make up. Truths about wholes might not be derivable from truths about their parts, but this does not mean that wholes are something in addition to the parts—in many cases, the parts massively interactively organised as they are. Nor does the apparent fact that talk of mental phenomena cannot be eliminated in favour of talk of physical phenomena entail that mental and physical phenomena are distinct. Taxonomic distinctness does not license an inference to metaphysical distinctness.

If you accept this is how it is, you can make sense of the place of the various sciences and their relative autonomy. You can also understand how, despite their autonomy they are unified at a deeper level: they describe and explain one and the same cosmos. One cosmos is cosmos enough.

References

Andersen, P. B., Emmeche, C., Finnemann, N. O., & Christiansen, P. V. (Eds.). (2000). *Downward causation: Minds, bodies, and matter*. Aarhaus University Press.

Davidson, D. (1967). Truth and meaning. *Synthese, 17*, 304–23. Reprinted in Davidson (1984): 17–36.

Davidson, D. (1970). Mental events. In Foster and Swanson (1970): 79–101. Reprinted in Davidson (1980a): 207–225; Heil (2003): 685–699.
Davidson, D. (1973). Radical interpretation. *Dialectica, 27*, 313–28. Reprinted Davidson (1984): 125–39; and in Heil (2003): 286–97.
Davidson, D. (1980). *Essays on actions and events*. Clarendon Press.
Davidson, D. (1984). *Inquiries into truth and interpretation*. Clarendon Press.
Feigl, H., Scriven, M., & Maxwell, G. (1958). *Minnesota studies in the philosophy of science 2: Concepts, theories, and the mind–body problem*. University of Minnesota Press.
Fodor, J. A. (1981). The mind–body problem. *Scientific American, 244*, 114–23. Reprinted in Heil (2003): 168–182.
Foster, L., & Swanson, J. (Eds.). (1970). *Experience and theory*. University of Massachusetts Press.
Hare, R. M. (1952). *The language of morals*. Oxford University Press.
Hare, R. M. (1984). Supervenience. *Proceedings of the Aristotelian Society* (Supplementary Volume), *58*, 1–16.
Heil, J. (Ed.). (2003). *Philosophy of mind: A guide and anthology*. Oxford University Press.
Heil, J. (2021). *Appearance in reality*. Clarendon Press.
Kim, J. (1990). Supervenience as a philosophical concept. *Metaphilosophy, 12*, 1–27. Reprinted in Kim (1993): 131–60.
Kim, J. (1993). *Supervenience and mind: Selected philosophical essays*. Cambridge University Press.
Kim, J. (2000). Making sense of downward causation. In Andersen et al. (2000): 305–21.
Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. In Feigl, Scriven, and Maxwell (1958): 3–36.
Place, U. T. (1956). Is consciousness a brain process? *The British Journal of Psychology, 47*, 44–50.
Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, mind, and religion* (pp. 37–48). University of Pittsburgh Press. Reprinted as 'The Nature of Mental States' in Putnam (1975): 429–40. Reprinted in Heil (2003): 158–67.
Putnam, H. (1975). *Mind, language, and reality: Philosophical papers* (Vol. 2). Cambridge University Press.
Rovelli, C. (2016). *Seven brief lessons on physics* (S. Carnell & E. Segre, Trans.). Riverhead Books.
Smart, J. J. C. (1959). Sensations and brain processes. *Philosophical Review, 68*, 141–56. Reprinted in Heil (2003): 116–27.
Sperry, R. W. (1969). A modified concept of consciousness. *Psychological Review, 76*, 532–536.
Strong, C. A. (1923). *A theory of knowledge*. Constable & Co. Ltd.

Chapter 8
Why Functionalism Is a Form of 'Token-Dualism'

Meir Hemmo and Orly Shenker

Abstract We present a novel reductive theory of type-identity physicalism (called Flat Physicalism), which is inspired by the foundations of statistical mechanics as a general theory of natural kinds. We show that all the claims mounted against type-identity physicalism in the literature don't apply to Flat Physicalism, and moreover that this reductive theory solves many of the problems faced by the various non-reductive approaches including functionalism. In particular, we show that Flat Physicalism can account for the (alleged) appearance of multiple realizability in the special sciences, and that it gives a novel account of the genuine autonomy of the kinds and laws in the special sciences. We further show that the thesis of genuine multiple realization, which is compatible with all forms of non-reductive approaches including functionalism, implies what we call token-dualism; namely the idea that *in every token* (that partakes in this multiple realization) there are non-physical facts, which may either be non-physical properties or some non-physical substance; that is, we prove that non-reductive kinds necessarily assume non-reductive tokens, i.e., *token-dualism*. We then consider a surprising feature of our approach, in which, despite its fully reductive nature, the special sciences are still genuinely autonomous, after all. Finally, we show that all forms of non-reductive approaches including functionalism imply a literally multi-leveled structure of reality.

M. Hemmo (✉)
Philosophy Department, University of Haifa, Haifa, Israel
e-mail: meir@research.haifa.ac.il

O. Shenker
Sidney M. Edelstein Center for the History & Philosophy of Science, Technology and Medicine
The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: orly.shenker@mail.huji.ac.il

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_8

8.1 The State of Art Concerning Reductive Type-Identity Physicalism

> Materialism as a theory of the nature of the world has had a curious history. Arising almost at the beginning of Greek philosophy, it has persisted down to our own time, in spite of the fact that very few eminent philosophers have advocated it. ... A system of thought which has such persistent vitality must be worth studying, in spite of the professional contempt which is poured on it by most professors of metaphysics. (Russell, 1925, p. V.)

These words of Bertrand Russell, written almost a hundred years ago, still describe the state of the art in philosophy, if the term 'materialism' is interpreted as *reductive type-identity physicalism* (*"reductive physicalism"* for short). In contemporary philosophy there are two major lines of thinking that reject reductive type-identity physicalism. One rejects physicalism of all sorts, and endorses dualism; for example, arguments around the so-called "hard problem of consciousness". We don't address these arguments here.[1],[2] The other line of thinking that rejects reductive type-identity physicalism consists of the variety of approaches all of which fall under the general title "non-reductive physicalism". Some central representatives of this line of thinking are our target in this paper. In contemporary literature, this second line of thinking is so dominant, that there is sometimes a tendency to disregard the alternative altogether. Here are two very recent examples. Elpidrou (2018), in "Introduction: The Character of Physicalism" to a special issue of the journal *Topoi*, begins by saying that "Not many issues in philosophy can be said to match, let alone rival, physicalism's importance, persistent influence, and divisiveness" (p. 435), and continues to characterize physicalism in such a way that "various identity theories will not be forms of physicalism as understood here." (p. 437). Tiehen (2018), in an overview of contemporary literature in the journal *Analysis*, titled "Physicalism", begins by characterizing "Physicalism" as the thesis that "*that there is nothing over and above the physical*", and continues to characterize the notion of "nothing over and above" in such a way that reductive physicalism is conspicuously not taken seriously.[3]

[1] We don't find arguments based on the so-called "hard problem" very convincing. For example, conceivability arguments, such as Kripke's (1980) or Chalmers's (1996), presuppose that the non-identity of mental kinds with physical kinds is *conceivable*. We reject this intuitive assumption. Since according to type-identity physicalism mental kinds *are* physical kinds, the option of non-identity (as in e.g., the "zombie scenario") is a *contradiction*, and therefore it is *inconceivable* in any interesting way (and not only metaphysically impossible!). But as we said, this is not our topic.

[2] We set aside here Hempel's dilemma, because: (i) it applies to all forms of physicalism while this paper focuses on the distinction between reductive and non-reductive physicalism; and (ii) it applies not only to physicalism but equally to dualism (or any other deep structure theory). For our take on Hempel's dilemma, see (Firt et al., 2021).

[3] In (Brown & Ladyman, 2019), which provides a historical and philosophical overview of materialism and physicalism, the section on "Mind-Brain identity theory" (in Ch. 6) consists of (a) a non-critical summary of the introductory essay of a collection of essays from 1969, and (b) a non-critical presentation of a quotation from an essay by Crane in the *Time Literary Supplement*

Of course, there are arguments against reductive physicalism (for example, the argument that the mental is computational); but unless this view is presented, taken seriously, and (even!) defended, the game isn't over; and, as we show in this paper, it certainly isn't. Reductive type-identity physicalism should resume a position at the center of the philosophical stage, as an important detailed view that has a lot to be said in its favor, and which is – as we will show – much stronger than the varieties of non-reductive physicalism in solving some of the major problems faced by contemporary philosophy of mind and of science. Reductive type-identity physicalism is at one end of the spectrum of ontological theories (we will not try to characterize the other end of this spectrum); and understanding it is necessary in order to obtain a good understanding of the entire spectrum. Indeed, in this paper we shall point out some results of clarifying this point.

One reason why reductive physicalism was rejected by so many thinkers in the past, and is still rejected (or ignored) by the majority in the present, might be that the theories of reductive physicalism that have been described in the literature are, simply, not good. While people have made a lot of efforts to develop non-reductive theories, and there is a variety of them in the literature, that incorporate a variety of metaphysical assumptions and respond to a variety of objections, the reductive theories that are presented aren't very well developed. Here are two examples that, while not very recent, are still influential.

One very influential description of a reductive type-identity physicalist theory is by Smart (1959),[4] who writes: "It seems to me that science is increasingly giving us a viewpoint whereby organisms are able to be seen as physico-chemical mechanisms. ... That everything should be explicable in terms of physics except the occurrence of sensations seems to me to be frankly unbelievable." (p. 142) Importantly, Smart says explicitly (on p. 143) that "the above is largely a confession of faith", not an argument. To see how Smart characterizes the physicalist view, consider his characterization of dualism: "There does seem to be, so far as science is concerned, nothing in the world but increasingly complex arrangements of physical constituents. All except for one place: in consciousness." (p. 142) Reductive physicalism simply omits the second sentence. This is a starting point for constructing a physicalist theory, but certainly calls for more details and for more arguments; without them it is quite vague.

Putnam (1975) adds some details to this very general characterization of reductive physicalism. In the paper in which he proposed the computational functionalist theory of mind,[5] Putnam criticized the "brain-state hypothesis" or the "physical-chemical hypothesis." (Sect. 8.3). What exactly is that hypothesis, according to this suggestion? Concerning the computational functionalist hypothesis, that is proposed

(2017) that expresses support of the "irreducible reality of the mental" (an essay that ends with the sentence "We will make no progress at all until we move beyond the simplistic brain's eye view.").

[4] It has been suggested that Smart should be read as a non-reductionist. Be that as it may, our interest here is in his characterization of the reductionist view; see also (Smart, 2017).

[5] See (Shagrir, 2005) for a retrospective exposition of the computational approach to the mind.

in that paper, Putnam writes: "This hypothesis is admittedly vague, though surely no vaguer than the brain-state hypothesis in its present form" (p. 434), and a bit later, "I contend, in passing, that this hypothesis, in spite of its admitted vagueness, is far less vague than the 'physical-chemical state' hypothesis is today, and far more susceptible to investigation of both a mathematical and an empirical kind." (1975, p. 435) Of course, both computational functionalism and the physical-chemical state theories were in their scientific infancy at the time, and to a large extent they still are. What, in particular, is the theory that he criticizes as too vague? The characteristics of this theory that Putnam provides are the following. (i) It is about the physical-chemical state of the brain. (ii) It is incompatible with psychophysical dualism (unlike functionalism: Putnam writes that "the functional-state hypothesis is not incompatible with dualism!" (1975, p. 436) (iii) Since the third characteristic has been immensely influential on arguments against reductive physicalism, let us bring it in Putnam's words (despite its length).

> Consider what the brain-state theorist has to do to make good his claims. He has to specify a physical-chemical state such that any organism (not just a mammal) is in pain if and only if (a) it possesses a brain of a suitable physical-chemical structure; and (b) its brain is in that physical-chemical state. This means that the physical-chemical state in question must be a possible state of a mammalian brain, a reptilian brain, a mollusc's brain (octopuses are mollusca, and certainly feel pain), etc. At the same time, it must not be a possible (physically possible) state of the brain of any physically possible creature that cannot feel pain. Even if such a state can be found, it must be nomologically certain that it will also be a state of the brain of any extra- terrestrial life that may be found that will be capable of feeling pain before we can even entertain the supposition that it may *be* pain. It is not altogether impossible that such a state will be found. Even though octopus and mammal are examples of parallel (rather than sequential) evolution, for example, virtually identical structures (physically speaking) have evolved in the eye of the octopus and in the eye of the mammal, notwithstanding the fact that this organ has evolved from different kinds of cells in the two cases. Thus it is at least possible that parallel evolution, all over the universe, might *always* lead to *one and the same* physical "correlate" of pain. But this is certainly an ambitious hypothesis. Finally, the hypothesis becomes still more ambitious when we realize that the brain state theorist is not just saying that *pain* is a brain state; he is, of course, concerned to maintain that *every* psychological state is a brain state. Thus if we can find even one psychological predicate which can clearly be applied to both a mammal and an octopus (say "hungry"), but whose physical-chemical "correlate" is different in the two cases, the brain-state theory has collapsed. It seems to me overwhelmingly probable that we can do this. Granted, in such a case the brain-state theorist can save himself by ad hoc assumptions (e.g., defining the disjunction of two states to be a single "physical-chemical state"), but this does not have to be taken seriously. (Putnam, 1975 pp. 436).

Putnam's account is certainly more detailed than Smart's, in that he puts forward conditions that a reductive physicalist view must satisfy. However, it seems that Putnam has in mind a very vague idea of a "brain state" theory, and it may be due to the poverty of this theory that he takes it to be unreasonable in the way that he does. Indeed, we present in this paper a much more detailed theory of reductive type-identity physicalism, and the reader may return to the quotation from Putnam

later on and see that what Putnam sees as an unreasonable ambition is, in fact, an extremely successful line of *actual* scientific research.[6]

8.2 The Tasks of Flat Physicalism

Our aims in this paper are as follows. (1) We wish to present a new version of reductive type-identity physicalism, which we call "Flat Physicalism". We will show that this theory has the resources to respond to traditional objections to reductive physicalism. In particular, it's a type-type identity theory according to which there is no genuine multiple realization. (2) We will show that non-reductive physicalism is really a form of dualism, i.e. what we call token-dualism on which every token of a non-reductive kind has some non-physical aspect (property or substance). (3) We will show that on Flat Physicalism there are no "levels of reality". Before we start, let us explain these tasks in a bit more detail.

8.2.1 Task I: Constructing the Flat Physicalism Theory of Reductive Type-Identity Physicalism

In view of this state of art, our first task is to propose a *new* theory of reductive type-identity physicalism, that we call "*Flat Physicalism*." This theory is based on recent results in the philosophy of physics, specifically the philosophy of statistical mechanics, and it takes statistical mechanics as a paradigmatic example on the basis of which the notion of "physical kinds" is to be understood.

Our next tasks are to explain how this theory responds to some major criticisms mounted against reductive type-identity physicalism (e.g., the worries expressed by Putnam in the above quotation), and show that it solves problems faced by varieties of so-called non-reductive physicalism in ways that are much better than those of the latter.

For example, we will show how Flat Physicalism accounts for the appearance of multiple-realizability in the special sciences; explain why the (so-called "high-level") regularities described by special sciences (for example probabilistic regularities, or even irregularities), seem to be independent of the regularities described by the (so-called "lower-level") laws of physics, and consequently, in what sense the laws of the special sciences may be genuinely autonomous.

[6] It seems to us that this conclusion is also supported by Polger and Shapiro (2016), even though we don't endorse their characterization of realization and of the special sciences' kinds that comes with it.

8.2.2 Task II: Provide a Flat Physicalist Account of the (Alleged) Appearance of Multiple Realization

As is well-known, many forms of so-called non-reductive physicalist approaches allow for multiple-realization of the kinds that appear in the special sciences by physical kinds. Whether or not such multiple realization is observed in experience is an empirical question that we don't address here (see e.g. Bickle, 2010; Polger & Shapiro, 2016). Our task is to prove that genuine multiple realizability entails psychophysical dualism (either property or substance); and so *if* genuine multiple realizability obtains in the world, *then* psychophysical dualism (either property or substance) obtains in the world.

The idea of multiple-realization has been introduced (in different terms) as part of a non-reductive approach by Putnam (1975) in his proposal for a computational-functionalist theory of the mind, and developed by (e.g.) Fodor (1974), who took it to be one of the salient motivations for developing non-reductive approaches to the special science. However, as we will show, genuine multiple-realization of special science kinds by physical kinds entails what we call *token-dualism*, that is: it entails that *in every token* (that partakes in this multiple realization) there are non-physical facts, which may either be non-physical properties or some non-physical substance. It is sometimes said that in (so-called) non-reductive physicalism each token is physical and yet the special sciences' kinds aren't reducible to physics; this idea is sometimes called "*token physicalism*". We shall prove that non-reductive kinds necessarily assume non-reductive tokens, i.e., *token-dualism*. We show that this is the case even if the special sciences' kinds supervene on physical kinds: in this sense, *supervenience isn't sufficient* to ensure that the world is physical. The only theory in which the tokens are physical is that of a reductive type-identity physicalism that doesn't allow for multiple-realization.[7]

8.2.3 Task III: Show That If Flat Physicalism Obtains, Then There Are No Levels of Reality

The third task of this paper is to explain in what sense the ontology in our proposed type-identity theory is flat, that is, that there *are no* levels of reality; hence the name "Flat Physicalism". Whether or not one chooses to use the term "levels" to describe certain features of reality, so as to obtain so-called "levels of explanation" or "levels of descriptions" etc., is immaterial to this ontological claim. As a corollary we will show that any approach which is compatible with multiple-realizability, regardless

[7] Our argument (in Sects. 8.4 and 8.5) that multiple-realizability implies token-dualism is independent of other arguments against non-reductive approaches.

of whether it also assumes supervenience of high-level kinds on low-level kinds, entails multiple-levels of reality (not only of explanations, descriptions, etc.).[8]

8.2.4 *Task IV: Show That Flat Physicalism Allows for an Autonomy of the Special Sciences and Is Compatible with all Forms of Special Sciences Laws, Including Probabilistic Ones and Even with Cases of Special Sciences Anomaly*

Fodor (1974) famously emphasized that one argument in favor of a non-reductive picture is the fact that the special sciences progress independently of physics, and the forms of their laws appear to be very distinct from that of the laws of physics; they are *autonomous*, in this sense. Our task is to show that these observations are completely compatible with Flat Physicalism, which can account for them. We will explain in what sense the special sciences are autonomous within reductive type-identity physicalism, and will explain how special sciences laws, that have forms very different from those of physics, can come about. One well known example is the time-directed second law of thermodynamics that obtains phenomenologically despite the temporal symmetry of fundamental physics; we shall not address this case in detail here (see e.g., Frigg, 2008; Hemmo & Shenker, 2012, 2016; Sklar, 1993; Uffink, 2007). We will show, in general outline, how special sciences laws that are probabilistic, and even cases of anomalous phenomena (e.g., the anomaly of the mental conjectured by Davidson, 1970) can come about within a Flat Physicalist picture (see also our 2021a).

The paper is structured as follows. In Sect. 8.3 we present our proposal called Flat Physicalism which is a full-fledged reductive type-identity theory *inspired* by recent results in the philosophy of physics, mainly the philosophy of statistical mechanics. In Sect. 8.4 we consider in detail the strict identity of physical kinds and the special sciences' kinds, and show that our physical theory is rich enough to explain the appearance of multiple-realizability in our experience in terms of physical kinds that are *identical* to the special science kinds. In Sect. 8.5 we show that any view which is compatible with genuine multiple-realizability (such as all forms of functionalism including computational and causal[9] functionalism) entails that there is something *non-physical*, which is *present* in each and every (*token*-)occurrence of the multiply-realizable kind. In Sect. 8.6 we show explicitly on the basis of Flat Physicalism how the special sciences could be autonomous and even anomalous in the sense that the (physical) kinds they describe don't strictly satisfy any law. In Sect. 8.7 we show that

[8] See Bechtel (2016) for other arguments as to why the mechanistic multi-leveled ontology should be flattened. But he retains levels of *explanation*.

[9] Whatever the notion of causation is; see (Ben-Menahem, 2018; Frisch, 2015) for different notions of causation and their relation to physics.

non-reductive approaches are committed to a picture of reality which is *literally* of multiple levels. Section 8.8 is the conclusion.

8.3 Constructing Flat Physicalism: A Novel Theory of Reductive Type-Identity Physicalism

As we said above, our first task in this paper is to present a reductive physicalist theory, in which the special sciences' kinds are physical kinds by *strict iden*tity. The details of our approach are inspired by recent results in the philosophy of physics, mainly the philosophy of statistical mechanics (see Hemmo & Shenker, 2012, 2013, 2015a, 2016; Shenker, 2017a, b; see Shenker, 2018 for the quantum case), but in order to follow the discussion here, one need *not* be acquainted with the details of that theory.[10] We take statistical mechanics to be *a general theory of physical kinds*, and accordingly, will present its principles in the most general way. There are, in the literature, ample discussions of *special science kinds*, and of the ways in which they can relate to the physical kinds; there is much less discussion of what a *physical kind* is, and since such an analysis is crucial for understanding reductive identity physicalism, we undertake it in this section.

The main idea of Flat Physicalism is just this: the world is as described by physics, and this is *everything* that there is. Nothing is left out. The starting point is the *tokens*: by assumption, every token state of affairs is *fully* described by physics.[11] Understanding the tokens is key to our ideas; types will come out of them as we show below. We stress that by saying that everything is physical we mean that everything is identical to something physical, that is we have in mind a strict identity theory (that is, as we shall see, a type-identity theory); in particular we do not mean the weaker claim that everything supervenes on the physical (or some other metaphysical relation of dependence, such as grounding, realization, etc.). We prove in Sect. 8.5 that if the relationship between physical kinds and special sciences' kinds satisfies supervenience (of any sort) but does *not* satisfy type-identity, then (what we call) *token-dualism follows.*

Let us illustrate our ideas with an example. Our toy model will be a universe that can be coherently, fully and correctly described by classical mechanics.[12] In

[10] For the standard approaches to the foundations of classical statistical mechanics, see e.g., (Frigg, 2008; Sklar, 1993; Uffink, 2007). For our approach, see (Hemmo & Shenker, 2021a, b, c, d, 2016, 2019a, b, c; Shenker 2017a, b) and for the quantum case (Shenker, 2018).

[11] Scientific realism doesn't imply this characterization of a token. "Realism is about what is real and not about what is fundamentally real." (Psillos, 2009, p. 38).

[12] We agree with Ladyman and Ross (2007) and Wallace (2001) that it is a mistake to carry out metaphysical investigations, assuming that classical mechanics is unrestrictedly true. At the same time, the use of classical terminology and laws is legitimate if they preserve essential features of the phenomena and fundamental facts being addressed, which is the case here, as it brings out the main ideas of Flat Physicalism.

classical mechanics one can distinguish between two kinds of tokens: *token-states* and *token-sequences*. In classical mechanics a token state – called a *microstate* – consists of the exact positions and velocities of all the particles in the universe at a point of time[13]; and the corresponding example of a token-sequence is a sequence of such microstates, generated by the equations of motion, and often called a *micro-trajectory*. According to classical mechanics, given a microstate of the universe, together with parameters such as the mass of each particle, constraints such as the volume available to them, and limitations such as the total energy of the universe, the equations of motion (ideally) yield a continuous infinite temporal sequence of microstates. The connection and distinction between token-states and token-sequences, illustrated by the distinction between microstates and micro-trajectories, is significant for our analysis of functionalism (below).

Suppose that we are given the full details of some physical token-state or token-sequence; and suppose, as assumed by Flat Physicalism, that physics is complete, so that the physical microstate or trajectory describes everything that there is. And then suppose that we want to talk about some special science, for example: we want to talk about some laws of biology or of geology. Many think that the terms, the properties and laws of physics don't capture those of these special sciences, and therefore in order to describe these features of the universe we need to *add* something to the physical microstate or trajectory, that is, add something which isn't part of what we took to be the complete (physical) description of what there is. The complete physical description of the world, so they say, *misses out* something. This claim is *rejected* by Flat Physicalism, that offers the following alternative.

According to Flat Physicalism, the microstate of the universe is everything that there is, so that one cannot say *more* about the special science kinds beyond specifying this microstate. The only thing that remains, if one wants to say something *different* from physics, is to say *less* than physics does. And this is the route taken by Flat Physicalism: when we talk about special science kinds we say l*ess not more*: we refer to an *aspect* of the token-state or token-sequence (that is, respectively, an aspect of a mechanical microstate of the world or of its micro-trajectory), and this aspect is given to us by a *partial description* of the token. So, according to Flat Physicalism the special sciences are about certain aspects of the physical tokens of the world, and *they cannot be anything else because there isn't anything else*.[14]

[13] Of course, relativistic considerations should enter at this point, in non-classical physics. In standard quantum mechanics the microstate is the *pure* quantum state in Hilbert space; see (Shenker, 2018) for our view concerning the foundations of quantum statistical mechanics.

[14] An anonymous referee argued as follows. "Non-reductive physicalism. Like Flat Physicalism, agrees that the microstate of the universe is everything that there is (that's the "physicalism"!). But it acknowledges that there are many ways of *categorizing* microstates. One could in principle take the set of microstates that occur in London on Jan first 2020 and put them in a category. Or the set of microstates that were ever contemplated by my grandmother, and put them in a category. And so on. To describe these categories, we say "less not more", just as Flat Physicalism says: we refer to just one aspect of any token microstate, i.e., that it was contemplated by my grandmother. In effect,

A well-known example of an aspect in our sense is the identity statement "heat *is* molecular motion", or: "the temperature of an ideal gas in equilibrium *is* the average kinetic energy of its particles". In classical mechanics (which is our example), average kinetic energy of the particles of a sample of an ideal gas is only an aspect, given by a partial description, of a microstate of the universe, or more specifically, of the gas in question, which is a subsystem of the universe. There are other aspects (and other details about) the microstate of the universe at the moment of interest, and even other details of that subsystem, that aren't given by this partial description; for example, the positions of the gas particles, their total number, their specific velocities. Any such aspect is part of, or is literally in, the microstate: it is there in the same sense that the entire microstate is there (with the aspects on which we don't focus), even if we aren't looking at it, as it were. Once one (for example, Laplace's proverbial Demon) has access to the details of the microstate (or the micro-trajectory) of the universe, one can *derive* from it any aspect, by *ignoring* some details of the microstate or micro-trajectory (see also Portides, 2019). In that sense, there is nothing above and beyond the microstate (or micro-trajectory) described by physics; this is all that there is.

Flat Physicalism is an *identity theory* that is *not eliminativist*,[15] in the following sense. According to Flat Physicalism, the facts that are described by the special sciences are out there, as it were (subject to all the arguments for and against scientific realism). Science tells us that when referring to them we actually refer to the appropriate aspect of the fundamental physical microstate. Water *just is* H_2O (see Chang, 2012; Hoefer & Marti, 2019); it isn't correlated with H_2O. Similarly, to use our example, temperature (of an ideal gas in equilibrium) is (identical to) average kinetic energy. It would be a *mistake* to say in statistical mechanics that, for example, average kinetic energy "gives rise to" temperature or "grounds" temperature, since according to statistical mechanics, there are no facts (any sort of facts) in the world beyond the actual microstate of the universe, its aspects and its sequence over time. In statistical mechanics there are no "mereological facts" which are about "combining Lego parts" as it were (see Chang, 2012; and our 2021a), since a *conceptually and physically inseparable* part of the physical description is the

that's a partial description of a set of microstates – partial, because it does not completely describe any one of them. In that sense it says "less not more". The non-reductive physicalist can agree with all this. Their key point, as I understand them, is that the categories of the special sciences don't line up neatly with the categories of physics; e.g., just think of the set of microstates that were contemplated by my grandmother: they needn't have anything in common *physically* speaking; from the perspective of physics they may appear a heterogeneous mix. As I understand them, non-reductive physicalists think that special science categories are like *that*. But that is consistent with the idea that to say something different from physics is to say less not more." We shall see later that: (i) there is a straightforward type-identity account of "the set of microstates that were ever contemplated by my grandmother", namely the macrovariable pertaining to my grandmother's contemplating! See Sect. 8.4; and (ii) if the categories (i.e., kinds) of the special sciences turn out to correspond to sets of microstates that don't have a similar account, then token-dualism follows; see Sections 8.4 and 8.5.

[15] For an overview of eliminativism, see (Ramsy, 2019).

intermolecular interactions. In this sense, statistical mechanics (when successful) accounts for the thermodynamic phenomena in term of strict *identity statements* of the kind 'temperature is molecular motion.'

As Smart wrote already in (1956), "you cannot correlate something with itself. You correlate footprints with burglars, but not Bill Sykes the burglar with Bill Sykes the burglar." (The same point is made by Papineau, 1993.) The discovery that water is H2O doesn't eliminate the notion of water, but enriches it.

A Question Arises Every token (state or sequence) has infinitely many aspects, each being a function of the complete microstate of the universe. All of them "exist" with the full token. Nevertheless, only a relatively small number of aspects appear in our experience and in our theories. Why is that so, and how the "preferred" (as it were) aspects are selected? Here is the Flat Physicalist explanation for this.

Consider two interacting systems (each is a sub-system of the universe), where the interaction between them is such that certain aspects of the microstate of each of them become correlated with certain aspects of the microstate of the other.[16] Although both are completely and fully on a par as being physical systems, and although physics is blind to any roles we might ascribe to any of the systems, it is convenient, for the purpose of our illustration, to call one of them a 'measuring device', and the other a 'measured system', and we shall say that the measuring device is *sensitive* to certain aspects of the measured system, and not to others, in the case that from the end state of certain aspects of the measuring device, one can tell which was the value of the corresponding aspects of the measured system to which the measuring device is sensitive, at a certain time of interest, say the time of interaction. Our sense organs are such measuring devices, and (presumably) so are our brains: and they are sensitive to certain aspects of our environment (and not to others). Those aspects of our environment to which we are sensitive appear in the special sciences. If we are lucky (and we sometimes are) the aspects to which we are sensitive satisfy certain regularities, a fact that we can use to make predictions (and evolutionary explanations can be given for this fact). It is a task of the special sciences to identify the aspects that satisfy useful regularities and to formulate these regularities. Another task of the special sciences is this. We can extend the set of the aspects of our environment to which we are sensitive by building new measuring instruments, following scientific discoveries concerning regularities governing certain additional aspects of our environment, to which our sense organs were not initially sensitive. The instruments are sensitive to these additional aspects of the environment, to which we have no direct physical access; and to read these instruments we employ aspects of them to which we are sensitive. Identifying such further aspects of the environment that satisfy regularities, hitherto

[16] More precisely, the microstate of the universe evolves along its micro-trajectory such that if we focus on the two sets of the degrees of freedom associated with these two systems, and of certain aspects of the microstates of them, then we find that those aspects are correlated. We make here the assumption that the microstate of the universe is separable, which is standard in classical physics though not in quantum mechanics.

unobserved by our sense organs, and which may be interesting and useful, is also the job of the special sciences. In this process of discovery and selection the special sciences are *autonomous* (see [author reference]); more on the autonomous status of the special science kinds and laws in Sect. 8.7). It turns out, then, that the selection of "preferred" aspects of the environment is relative to an observer, possibly aided by measuring instruments; and the special sciences identify them and their regularities. This role of the special sciences, which is of extreme importance in understanding our world, explains their autonomy; we return to this point in Sect. 8.7.

Consider, now, a case in which a measurement interaction is carried out between a measuring device and a certain system in the environment, by the end of which a certain aspect of the state of the device is correlated with a certain aspect of the measured system. For convenience of presentation, suppose that the measuring device is *us*, and our brain state by the end of the interaction registers the aspect of the environment to which we are sensitive. At this point, *all* we know (following this interaction) about the microstate of our environment is that it has this aspect; we know nothing about the other (infinitely many) aspects of the environment. The important point is that there can be many *counterfactual* microstates that *share* this (known) aspect, but differ in the other (unknown) aspects; and they would all look to us the same as the actual microstate with which we have interacted. A *set* of microstates that *share* an aspect is called a *macrostate*. Corresponding to the two sorts of tokens: token-states and token-sequences, there are two sorts of macrostates: macrostates that are *sets of microstates* that share the same physical aspect, and macro-sequences that are *sets of trajectories,* or of sequences, that share suitable physical aspects.[17] (The latter will be important in understanding what a "function" is and what "functionalism" is in Flat Physicalism.)

The fact that two microstates of the universe (or of some sub-system of it) share the same aspect, and therefore belong to the same macrostate, is a fact about these two microstates that obtains regardless of whether or not there is a measuring device that is sensitive to it or interacts with it. At the same time, the fact that we – as measuring devices – are sensitive to certain aspects of our environment and not to others, explains the apparent preferred status that certain aspects of the environment have, in that they feature in our experience. To repeat and to emphasize, it is a conceptually distinct matter whether these aspects of our environment satisfy some regularity. However, it makes sense to assume that there would be an evolutionary advantage to creatures that are sensitive to aspects of the environment that also satisfy regularities, thus enabling predictions.

With respect to macrostates it is important to notice the following fact. Consider the example "temperature is molecular motion". The "molecular motion" in question is an aspect of the microstate of a sample of gas. When we carry out a measurement interaction with a particular sample of gas, in order to measure the

[17] There are various notions of kinds and properties that come up in such discussions. Our arguments here don't depend on the details of these analyses, as long as the latter are compatible with physics.

temperature of that gas at that event, we in fact interact with the suitable aspect (namely, the "molecular motion") of the *particular* microstate that obtains at that *particular* moment. We emphasize that we *never* measure a *set* of microstates, since the set consists of the *actual* microstate as well as (infinitely) many *counterfactual* ones; and in the measurement we interact with *actual* matters of fact, not counterfactual ones, that don't obtain (and most of which will *never* obtain) in the universe. Indeed, when we look at the *particular* event of the *particular* sample of gas, we *directly* know (by measurement) what its temperature is: We don't need to know which *other* (counterfactual) microstates would give rise to the same temperature (for all we know, or sense, or measure, this could be the only member of the set). Sometimes people say that what we actually measure are "macrostates". If by "macrostate" they refer to an *aspect* of the *particular actual* microstate (also sometimes called "macrovariable"), then that is correct; but if the claim is that we measure *sets* of microstates, then this statement is *wrong*.[18]

In Flat Physicalism, then, a physical token belongs to a physical kind if that token has the suitable aspect, ideally given by a partial description of it; and all the special sciences' kinds are physical kinds in this sense, just because the microstate and the micro-trajectory of the universe are everything that there is, *tout court*. But this point involves some further details, to which we now turn.

8.4 Special Sciences' Kinds in Flat Physicalism

Here is the *central* question concerning reductive versus non-reductive physicalism. Consider three tokens (or microstates) A, B and C, such that A and B belong to the same special sciences' kind M but C doesn't. *What is the fact in virtue of which this is the case?* Which facts in the world fix the partition of tokens into types? From now on we focus on this question, answering it in the framework of Flat Physicalism, comparing this answer to those of other frameworks, and studying its implications.

According to Flat Physicalism the fact in the world that fixes the partition of tokens (or microstates of some subsystem of the world) to types must be in the microstate (or micro-trajectory) *of the universe*, just because this is everything that there is. There are three options here:

(I) The fact that makes A and B (but not C) members of the same set (i.e., kind) is in A and B (but not C); and it is non-disjunctive.

[18] Above we gave an example of an aspect, namely, the average kinetic energy of the molecules of an ideal gas in equilibrium. Some people say that "averages" are above and beyond the physical facts, since, possibly, most of the sets of molecules don't have this particular velocity, and possibly, none has. But when we have the full details of the microstate, we already have this average; it can be logically derived from the microstate, and in this sense, it is already *in* the microstate. We don't need to postulate any additional facts over and above the microstate, to derive this fact. It exists, our there, as it were, even if nobody is interested in calculating it and will never actually derive it.

(II) The fact that makes A and B (but not C) members of the same set is in A and B (but not C); this fact is unavoidably disjunctive (see below).
(III) The fact that makes A and B (but not C) members of the same set is elsewhere in the microstate of the universe.

Let us explain a bit this classification of options. The reason that it emphasises the distinction between disjunctive (Option (I)) vs. non-disjunctive properties (Option (II)) is that.

our main targets in this paper are non-reductive views (including functionalism of all versions). As is well known, according to non-reductive views, special sciences' kinds are, as a matter of principle, *irreducible* to physical kinds. For example, in developing his theory of anomalous monism, Davidson (1970, p. 141) argued that "we can pick out each mental event using the physical vocabulary alone, but no purely physical predicate, no matter how complex, has, as a matter of law, the same extension as a mental predicate." Similarly, and in a way of supporting this particular point, Fodor (1997, p. 153) argued that: "it's exactly the distinction between disjunctiveness and disjunctive realization that functionalists are insisting on when they say that pain states are nomologically homogeneous under their functional description despite the physical heterogeneity of their realizers. (Fodor, 1997, p. 153), where one of the central motivations for holding this idea is the thesis of multiple realization (or realizability), which we discuss in detail below, especially with respect to the question of whether or not this thesis is consistent with physicalism. Of course, there are also other (related) motivations for holding non-reductive views, such as the question of how freedom (e.g. of choice) is possible in a world that's governed by natural laws (see e.g. Davidson[19]). But we set these issues aside in this paper.

Option (I):

Tokens A and B share a non-disjunctive aspect (or property; we use these terms interchangeably)) which is M, while C doesn't have *this* aspect. We, as measuring devices, are sensitive to this shared non-disjunctive property and register it.

[19] Here are two quotes of Kant on freedom and natural law that Davidson (1970, pp. 137–149) cites at the beginning and end (respectively) of his paper Mental Events: "[I]t is as impossible for the subtlest philosophy as for the commonest reasoning to argue heedom away. Philosophy must therefore assume that no true contradiction will be found between heedom and natural necessity in the same human actions, for it cannot give up the idea of nature any more than that of heedom. Hence even if we should never be able to conceive how freedom is possible, at least this apparent contradiction must be convincingly eradicated. For if the thought of freedom contradicts itself or nature... it would have to be surrendered in competition with natural necessity." And: "It is an indispensable problem of speculative philosophy to show that its illusion respecting the contradiction rests on this, that we think of man in a different sense and relation when we call him free, and when we regard him as subject to the laws of nature... It must therefore show that not only can both of these very well co-exist, but that both must be thought as necessarily united in the same subject."

An important result of this way of understanding special sciences' kinds as physical kinds is that it provides an explanation for the following *fact*. In every particular event in which we encounter a particular token in our environment (say A) we are able to say, directly and immediately, whether or not A is M, without having to know anything about B, or about any other aspect pertaining to the microstate of the world (or to any subsystem of it), and regardless of whether or not B ever obtains in the world or whether or not there are other aspects pertaining to the world that may or may not obtain. We detect the aspect M in the actual token A that obtains by physically interacting with it.

As we already said, these facts concerning the aspects M of A and B and not-M of C obtains whether or not these microstates are being observed by a measuring device that is sensitive to that aspect, or not; although we are naturally interested in those aspects to which we, as measuring devices, are sensitive. The non-disjunctive aspect that A and B (but not C) share may be very complex, and perhaps we will never find which aspect it is by empirical scientific investigation; but by assumption, it is there, and it accounts for the fact that A and B are of the same kind, but C is not.

Here again our example is instructive. Today we are so used to saying that temperature (of an ideal gas in equilibrium) is average kinetic energy, that we tend to forget that this was a highly non-trivial scientific discovery: nothing in the pre-scientific notion of "temperature" prepared us for the scientific identification of temperature with this particular complex mechanical aspect. Possibly, the aspects that are identical with special science kinds in biology, for example, are even more complex and hard to discover. But since according to Flat Physicalism there is *nothing in the world* except the microstate, and regardless of the complexity and the prospect of discovery, it follows that this is nevertheless the case. Flat Physicalism is the idea that the reduction of thermodynamics to mechanics, by way of such commitments, is to be generalized to all the special sciences.[20]

Option (II):

Suppose, however, that the fact that makes A and B (but not C) members of the same kind pertains to a disjunctive aspect (or property) of A and B, where the disjunction is unavoidable or indispensable in the sense that it cannot be replaced by a single predicate (however; this is Option (II)). This case is sometimes referred to in the literature as multiple realization (or realizability), which is characterized as follows.

> The multiple realizability thesis about the mental is that a given psychological kind (like pain) can be realized by many *distinct* physical kinds: brain states in the case of earthly mammals, electronic states in the case of properly programmed digital computers, green slime states in the case of extraterrestrials. (Bickle, 2013, our emphasis)

[20] Whether or not "temperature" is of case (I) or of case (II) is under debate. Compere (Frigg & Hoefer, 2015) and (List & Pivato, 2015) with (Hemmo & Shenker, 2019a).

In our terms, token A belongs (for example) to the brain state type and token B belongs to the computer type, where token C belongs to some other type (for example, it is of the type chair) which (we assume) is not a case of pain. What is meant by distinct physical kinds? On the one hand brain states and computer states are quantum states, and in this sense, they are not distinct. On the other hand, they are different quantum states, and this difference makes a difference in physical kind.[21] Here is what Fodor (1974) says about this sort of difference:

> I am willing to believe that physics is general in the sense that it implies that any event which consists of a monetary exchange (hence any event which falls under Gresham's law) has a true description in the vocabulary of physics and in virtue of which it falls under the laws of physics. But banal considerations suggest that a description which covers all such events must be *wildly disjunctive*. (Fodor, 1974, p. 103, our emphasis).

Fodor explains (in his 1997, p. 153 reply to Kim) this idea of physical descriptions that are wildly disjunctive:

> Though Kim says that he concedes that psychological properties are MR, that's only because he isn't distinguishing being MR (like pain) from being disjunctive (like jade). But it's exactly the distinction between disjunctiveness and disjunctive realization that functionalists are insisting on when they say that pain states are nomologically homogeneous under their functional description despite the physical heterogeneity of their realizers. (Fodor, 1997, p. 153)

So according to Fodor a psychological kind is multiply realizable just in case it can be realized by a wild disjunction of physical kinds, where by wild he (Fodor) means that the physical kinds in this disjunction (it is empirically discovered) are heterogeneous. That is: suppose that they don't share any *relevant* aspect, that can be associated with the special science kind (they may share (infinitely) many other aspects that aren't M). Many people, starting for example from Putnam (1975) and Fodor (1974) (albeit with important exceptions such as Polger & Shapiro, 2016) believe that this is indeed the case in our world. Whether or not this is the case, and whether or not the arguments for this are empirical, is debatable; but here we want to examine what Flat Physicalism has to say in case this is indeed true, and therefore we shall *assume* – at least to begin with – that there are sets of tokens that belong to the same special science kind but don't share any relevant physical aspect.

For supporters of multiple realization, such a case seems mysterious. Fodor famously wrote:

> 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 (not all the classes of things and events about which there are important, counterfactual supporting generalizations to make) are, or correspond to, physical natural kinds. (1974, p. 113)

> The very *existence* of the special sciences testifies to reliable macrolevel regularities that are realized by mechanisms whose physical substance is quite typically heterogeneous.

[21] Polger and Shapiro's (2016) propose to identify the similar and different aspects by asking experts. We are not committed to this view.

> Does anybody really doubt that mountains are made of all sorts of stuff? Does anybody really think that, since they are, generalizations about mountains-as-such won't continue to serve geology in good stead? Damn near everything we know about the world suggests that unimaginably complicated to-ings and fro-ings of bits and pieces at the extreme *microlevel* manage somehow to converge on stable *macro-level* properties. (Fodor, 1997, p.160)

An anonymous referee argued that even in cases of realization of a special science kind by wildly disjunctive physical kinds, as contemplated by non-reductivists, one may still hold a type-identity theory, as follows: "The core of non-reductive physicalism is that the categories of the special sciences don't line up with those of physics, i.e., the members of one special science category will be physically heterogeneous. So *if* the special science category is identical to a physical category, the latter will be massively disjunctive. In the language of physics, it may even be impossible to express the disjunctive category in a finite manner. But that's just a fact about language. It remains open for the non-reductive physicalist to say that the *properties* themselves (or kinds, or categories, whatever you want to call them) are identical. In this regard, recall that Fodor's main objection to reductive physicalism (in the 1974 paper cited) was not the bridge principles between special science properties and physical properties. It was that the physical properties involved in the bridge principles would be so massively heterogeneous that they can't be involved in physical laws. That's why he called the special sciences' laws "autonomous": if you take a special science law, and then replace the special science kinds with the physical kinds that the bridge principles associate them with, the result is *not* a physical law. That's why special sciences laws don't reduce to physical laws – non-reductive physicalism. But everything I just said is consistent with the special science kinds being identical with massively disjunctive physical kinds".

But what is the ontology in question? We take it that by "special science kinds being identical with massively disjunctive physical kinds", one means that the special science kinds are associated with wild disjunctions of physical kinds, so that they are realized (or realizable) by tokens belonging to this massively disjunctive set. Our question is this: which facts bring about special sciences' kinds, if the latter are "identical with massively disjunctive physical kinds"? Fodor famously continues in his "Conclusion (molto mysterioso)" of this paper, as follows.

> 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 . . . why there should be (how there could be) unless, at a minimum, macrolevel kinds are homogeneous in respect of their microlevel constitution. Which, however, functionalists in psychology, biology, geology and elsewhere, keep claiming that they typically aren't. So, then, why is there anything except physics? . . . 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, p. 161)

For supporters of Flat Physicalism, phenomena such as the ones mentioned by Fodor aren't mysterious at all, and can be fully explained in the framework of reductive physicalism, in the following way.

Consider three *mutually equally heterogeneous* tokens: A, B, and C. And suppose that A and B belong to the same special sciences' kind M, but C doesn't. Which fact makes it the case that this is the partition to the special science kinds? Why is B in the same set as A, but C – which is, by assumption, heterogeneous to the same degree! – isn't? The phrase "*to the same degree*" is very important here: B isn't "more similar" to A than C is, since if it were the case, A and B would share a physical coarse-grained aspect of their microstates, so that this second sort of special science kinds would collapse into the first sort. Therefore, we must preclude this case and suppose that C is as heterogeneous relative to A as B is; all three tokens are completely different from each other; they are all *mutually heterogeneous to the same degree*. This case is called, in contemporary literature, *multiple-realizability of special science kinds by physical kinds.* By assumption, the two following facts obtain: (a) tokens A and B (but not C) are of the same special sciences' kind; (b) tokens A, B, and C are equally mutually heterogeneous. How can (a) and (b) be reconciled?

An influential example of this idea was given by Fodor (1974) in order to illustrate his non-reductive approach. He writes:

> Gresham's law says something about what will happen in monetary exchanges under certain conditions. I am willing to believe that physics is general in the sense that it implies that any event which consists of a monetary exchange (hence any event which falls under Gresham's law) has a true description in the vocabulary of physics and in virtue of which it falls under the laws of physics. But **banal considerations** suggest that a description which covers all such events must be **wildly disjunctive**. Some monetary exchanges involve strings of wampum. Some involve dollar bills. And some involve signing one's name to a check. What are the chances that a disjunction of physical predicates which covers all these events (i.e., a disjunctive predicate which can form the right hand side of a bridge law of the form 'x is a monetary exchanged . . . ') expresses a physical natural kind? In particular, what are the chances that such a predicate forms the antecedent or consequent of some proper law of physics? The point is that monetary exchanges have interesting things in common; Gresham's law, if true, says what one of these interesting things is. But what is interesting about monetary exchanges is surely not their commonalities under physical description. A natural kind like a monetary exchange could turn out to be co-extensive with a physical natural kind; but if it did, that would be an accident on a cosmic scale. In fact, the situation for reductivism is still worse than the discussion thus far suggests. For, reductivism claims not only that all natural kinds are co-extensive with physical natural kinds, but that the co-extensions are nomologically necessary: bridge laws are laws. So, if Gresham's law is true, it follows that there is a (bridge) law of nature such that 'x is a monetary exchange iff x is P', where P is a term for a physical natural kind. But, **surely**, there is no such law. If there were, then P would have to cover not only all the systems of monetary exchange that there are, but also all the systems of monetary exchange that there could be; a law must succeed with the counterfactuals. What physical predicate is a candidate for 'P' in 'x is a nomologically possible monetary exchange iff Px'?
>
> To summarize: an immortal econophysicist might, when the whole show is over, find a predicate in physics that was, in brute fact, coextensive with 'is a monetary exchange'. If physics is general – if the ontological biases of reductivism are true – then there must be such a predicate. But (a) to paraphrase a remark Donald Davidson made in a slightly different context, nothing but brute enumeration could convince us of this brute co-extensivity, and (b) there would seem to be no chance at all that the physical predicate employed in stating the co-extensivity is a natural kind term, and (c) there is still less chance

> that the co-extension would be lawful (i.e., that it would hold not only for the nomologically possible world that turned out to be real, but for any nomologically possible world at all). (Fodor, 1974, pp. 103–104; our empheses)

We set aside for a moment the fact that Fodor's argument here is based on unspecified "*banal considerations*".[22] We shall comeback to Option (II) of disjunctive kinds below in this section. Let us focus now on the example itself which is this: things that are physically heterogeneous, for example, strings of wampum or dollar bills (today we would add states of certain electrical circuits) belong to the same economics kind of "monetary exchange". How can that be? Flat Physicalism offers the following answer, which brings us to Option (III), according to which the fact that makes microstates (say, A and B) be of the same kind M (i.e. the fact that makes A and B but not C be members of the same set) is not determined by the aspects (or properties of A, B and C, but rather by something elsewhere in the microstate of the universe.

Option (III):

To combine our abstract notations with Fodor's example, suppose that token A is a string of wampum, token B is a dollar bill, and token C is a chair (assuming that chairs aren't cases of monetary exchange; in Fodor's approach they *could* be (as a brute fact), but suppose that (as a brute fact) they *aren't*); and M is the kind "monetary exchange". Consider Fig. 8.1 in which the universe is partitioned into two sets of degrees of freedom: Along the vertical axis we depict the three systems with respect to which we ask whether their microstates are tokens of the kind M; and along the horizontal axis we depict the microstates of a system that we call The Device that starts out in some "ready state", and then according to the laws of physics evolves to a final pointer state that indicates "this is monetary exchange",

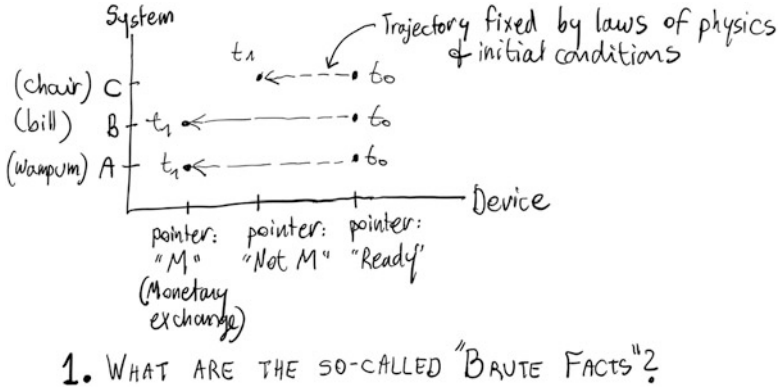

Fig. 8.1 What are the so-called "brute facts"?

[22] Since the entire argument is based on unexamined intuitions, we find it rather shocking that this paper (and its 1997 sequel) became so influential.

whereas if C obtains its final pointer state is "this wampums and dollar bills (but not chairs) are "monetary exchange", despite the fact that it doesn't measure (or otherwise reflect) any physical aspect of these tokens (which are equally mutually heterogeneous). The fact that The Device ends up in the same pointer state in cases A and B (but not C) isn't in virtue of anything physical about A and B, but it is, of course, in virtue of the physics of the total universe, that takes the total microstate of the universe from the initial microstate at t0 to the final microstate at t_1. The fact that this final microstate of the universe happens to be such that The Device points at M in cases A and B but not C seems "*brute*" in the sense that it cannot be explained by any aspect shared by A and B, which are completely heterogeneous; but this perspective is misleading: the full physical explanation of matters of facts in the universe is given by the complete micro-trajectory of the universe, and is revealed by proper partition of the degrees of freedom of the universe into subsystems. In our case, if we focus on the extended microstates of wampum+Device, dollar bill+Device, and chair+Device, then the special science kind "monetary exchange" turns out to be a feature shared by the first two extended microstates, that have the physical aspect M (but not with the third extended microstate, that has the aspect not-M). This fact is *fully explained* by fundamental physics, and *isn't brute* at all. In other words: the kind "monetary exchange" is *not* a feature of wampums and dollar bills that is "*measured*" (in some appropriate sense of the term) by The Device: The Device is *not a "measuring device"* for measuring the property "monetary exchange", and Fig. 8.1 does*n't* depict a measurement interaction; rather, "monetary exchange" is a feature of The Device itself, together with these other systems.

Constructing such a device isn't trivial, and the evolution described by the trajectories depicted in Figure1 may seem strange or even conspiratorial, since one cannot explain the motion of The Device by appealing to a measurement interaction that is sensitive to a shared aspect of A and B. But however strange this case is perfectly compatible with fundamental physics.

If indeed wampums and dollar bills don't share any physical aspect (as we assume here), then it turns out that nature herself has built an incredible device that operates as in Fig. 8.1, namely: us. Our brains evolve, following our interactions with society, in such a way that certain brain structures are what we call "conventions" and so on, in such a way that when we encounter a coin we *immediately and directly* enter the mental state of "entertaining the thought that this is monetary exchange". Regardless of what "conventions" (etc.,) are, in order for us to be able to recognize coins (or wampums, in the appropriate case) as monetary exchanges and act accordingly, in our daily lives, without constantly consulting with endless lists of tokens that might fall under this kind, our brains must be prepared in the physical (e.g., neuronal) state that assimilates these "conventions", and acts like The Device in Fig. 8.1.

Since the special science kind "monetary exchange" is a feature not of wampums or dollar bills, but of wampum+Device or dollar bill+Device, it is *not genuinely multiply realized* by physical kinds. It nevertheless gives *the impression* of multiple-realization, if we ignore The Device. *Ignoring* The Device makes our scenario a

case of *apparent, non-genuine*, multiple-realization: If we think of ourselves as The Device, then it is natural to expect that we should observe empirically the appearance of multiple-realization (as indeed many think that this is the case), since it is natural to treat *only* the external environment as "the system", ignoring the role of ourselves as The Device in the "extended system".

In Option (I) for constructing physical types, we stressed that the fact that two tokens share a non-disjunctive aspect is a feature of the world regardless of whether or not any measuring device is sensitive to this aspect and can register it. This is different from Option (III) of seemingly disjunctive partitions for constructing physical types: here, The Device is essential, since the physical kind is a feature of it. Lacking a Device, there is no non-disjunctive partition of the state space into the kinds M and not-M; there is no "monetary exchange", but only heterogeneous tokens that have nothing to do with each other.

An anonymous referee argued as follows. "There are two ways of developing theories like this. In one way, the property of being a monetary exchange is identified with the property of causing a certain reaction in The Device. On that view, being a monetary exchange is indeed a feature of The Device as much as the other systems and I see how this fits into Option (III). But here's another way to develop response-dependent views. Look at all those things (processes, events) that could possibly cause that reaction in The Device. For each one, take the physical property it has in virtue of which it caused that reaction. Each such property is "in" the thing, not The Device. Now disjoin all those properties. On this second approach, that disjunction is the property of being a monetary exchange. That disjunction doesn't involve The Device at all – The Device was just a "reference fixer", as it were, that allowed us to pick out the disjunctive property . . . I don't see why non-reductive physicalists couldn't endorse this approach; it amounts to a theory of why certain non-reductive kinds are "preferred" that a non-reductive physicalist could accept."

We do not reject this view on *a priori* grounds as a possible way of accounting for special sciences' kinds. But obviously this brings us back to the genuinely disjunctive case of Option (II), namely the case of genuine multiple realizability. So let's examine this idea a bit further and see what it implies.

Option (II),

the case of genuine multiple realizability (continued): Suppose, however, that one insists on the following ontological claim (as our anonymous reviewer just did): A, B and C are mutually equally heterogeneous, and A and B are M (but C isn't) even if there is no Device. Flat Physicalism has *no* resources to account for this case, and we are led to the conclusion that there is something non-physical in each and every token microstate, a fact that makes it belong to either one of these two sets, *a non-physical fact that we somehow perceive* in the presence of that token. We shall come back in more detail to this point in the next section.

8.5 Why "Non-reductive Physicalism" Is a Form of Token-Dualism

The case in which the physically heterogeneous tokens A and B are of the same special science kind M, where M may be, for example, a biological kind or an economics kind, can – as we have just seen – be accounted for in Flat Physicalism, by extending the token and bringing in The Device, whose aspects are shared by the A + Device token and the B + Device token. This route is, however, not available for psychological or mental kinds. Our main argument in this section is this. If The Device is us, and its pointer states "M" or "not M" are our psychological or mental states in which we entertain the thought "this is monetary exchange" or "this isn't monetary exchange" (or have some other corresponding mental state), then – according to Flat Physicalism – the set of (for example) "having M thoughts" cannot be multiply realized by heterogeneous physical kinds; all the "having M thoughts" cases must share a physical aspect, and the kind "has M thoughts" would be identical with the physical kind of having that physical aspect. This has the following consequence: The popular idea called "non-reductive physicalism" in which the tokens are physical but the kinds are (somehow) not reducible to physics, is *inconsistent*. *Either* psychological kinds *are* (reducible to, or identical to, or nothing over and above) physical kinds and tokens are physical, *or* (exclusive or) psychological kinds are *not* (reducible to, or identical to, or nothing over and above) physical kinds, and *token-dualism* obtains in the world. By "*token-dualism"* we mean that every token of the multiply-realizable kind contains a non-physical element, that may be a non-physical property of the token or some non-physical substance. This entails that *all* forms of functionalism (including computational functionalism), even if they require supervenience, as long as they allow for multiple realizability (even if they accept that in our world there is no multiple realization), are forms of token-dualism. *Supervenience is, then, not sufficient for physicalism, and is compatible with property or substance dualism*. Let us see why this is the case.

Suppose that the tokens A and B are microstates of two systems that share the same mental state. They could for example be the microstates of a human being and an octopus both of which feel the exact same kind of pain (to use Putnam's 1975 famous example; see discussion in Polger & Shapiro, 2016). And suppose that when The Device interacts with either A or B, it ends up in a microstate that has the physical aspect P, that indicates "being in pain". So far, the case is the same as the one of non-mental kinds (compare Fig. 8.1), as described in the previous section.

But here are the difficulties.

(i) If this case is indeed similar to the above one, in which the kinds are non-mental, then – as we have seen in Option (III) – the property P (here: being in pain) isn't a property of A or B (here: the human being or the octopus) but of A + Device or B + Device, or even a property of The Device. So if we say that you, the reader, are in pain, we don't speak about you, and there is no fact

of the matter concerning this mental state of yours, until some external Device is brought in to measure your brain state. This result seems to us unacceptable and contrary to the empirical starting point.

(ii) The alternative is to leave out The Device, and focus on A and B themselves. But A and B are – by assumption – completely genuinely heterogeneous. More generally, in such a case the set consisting of all physical microstates corresponding to the special science kind M is wildly (or better, genuinely) disjunctive. These microstates share nothing physical (short of their disjunction, if you wish). And if, in every token, the physical is everything that there is, then A and B share no relevant feature at all (short opf their disjunction), and don't form a non-disjunctive physical kind. In Fig. 8.1, recall, if we omit the horizonal axis (of The Device) the tokens on the vertical axis don't form (non-disjunctive) sets or kinds.

Let us remark that the idea of self-measurement will not work here, since it will collapse to either case (i) or to case (ii).

(iii) If, nevertheless (as friends of multiple realizability insist), the genuinely heterogeneous physical microstates A and B share some mental fact, namely: the fact of "being pain", that is: if *genuine multiple realizability* of the mental kind P by physical kinds obtains, then A and B must share a non-physical fact. Each of them has this non-physical fact, independently of whether or not the other ever obtains. Indeed, as we said, we feel pain directly in individual cases regardless of whether or not other pains, or what have you, obtain in the world. Thus, each *token* contains a *non-physical* element. It is immaterial for our argument whether this non-physical element is a property or a substance (and therefore we don't address this point here).

This is why, according to Flat Physicalism, the special sciences of psychology are radically different from the other special sciences (like biology or economics) in that the psychological kinds cannot even *apparently* be multiply realizable. We emphasize that this special nature of psychology has nothing to do with the so-called "hard problem" of consciousness or related issues.

In their defense of the idea of how high-level kinds can be multiply realizable by physical kinds, Davidson (1970), Fodor (1974, 1997), and Putnam (1975) seem to argue that multiple realizability is a consequence of the fact that the high-level special sciences sets (or kinds) are formed by "brute enumeration" (see Fodor, 1974 who follows Davidson on this point). Fordor (1974), for example, argues that it is a brute fact that tokens of heterogeneous physical kinds form a high-level set (or kind). Why should this be problematic? Our main point here is *not* that this idea by itself is logically incoherent; perhaps it is logically coherent (we don't take a stand on this question). Our point is that this idea is *incompatible* with another idea in physicalism, according to which the set of *all* possible tokens (e.g., the set of all possible microstates according to contemporary physics) is *all* that can possibly exist in the world, so that the sets (or the kinds) are determined completely by the tokens and nothing else, because there is *nothing* else. The "brute enumeration" (or facts) idea is that the sets are *not* determined by the tokens, but rather (somehow) by

the enumeration of their members. But this is misleading, since the critical question is: *how* these sets are formed, not how they are enumerated after they are formed (as it were). That is: given the set of all possible tokens, what in the *tokens* (or which facts about the tokens, which, again, are all there is) make it the case that some partitions into the sets obtain and some do not obtain. Sets determined by shared features of the tokens are formed by the facts about the tokens and nothing else. By contrast, sets formed by "brute enumeration" require *an enumerator*, that is, they require some *thing* or some *feature* or some (brute) facts external to (that is, over and above) the set of all possible tokens, which determine the enumeration, which in turn gives rise to the sets. And this (as we argued above) is token-dualism (substance or property) in disguise, since one has here *both* the tokens and the enumerator or the fact that determine that the enumeration is such-and-such.

If genuine multiple-realizability is a coherent idea (as many believe it is), then it could be true of psychological and cognitive states in our world; but then, importantly, mind-body dualism would be true of the world. In that sense, the empirical discovery of genuine multiple-realization would amount to an empirical discovery that psycho-physical dualism is true of the world.

8.6 Functionalism as Token-Dualism About Token-Sequences

Functionalism is sometimes taken to solve the main problem that we presented in Sect. 8.4 above. Recall: *The central question concerning reductive (versus non-reductive) physicalism* is this. Consider three tokens A, B and C, such that A and B belong to the same special science kind M but C doesn't. What is the fact in virtue of which this is the case? Which facts in the world fix the partition of tokens into types? Accepting that it may happen that A, B and C might be equally physically heterogeneous, so that the fact that A and B (but not C) are both M cannot be explained by a shared physical aspect, functionalists opted for the following explanation: A and B share the same "functional role". It may happen that two physical histories have the same functional structure, they emphasize, that is "realized" (this is a popular term in this context) in heterogeneous physical ways. Figure 8.2 illustrates this case: the states on the left hand side, indicated by numbers, are physically heterogeneous from the states on the right hand side, indicated by letters; but the structures are the same. Two tokens are deemed of the same "functional kind" if they occupy the same role in the functional structure, for example: State 1 and State A.[23]

We would like to very briefly note that in this sense functionalism doesn't solve the problem but repeats it. As we said above, there are two sorts of tokens: state tokens, and sequence tokens; and both are fully describable (ideally, of course)

[23] Whether the input and output states can be multiply realized is a subject of debate that we don't address here.

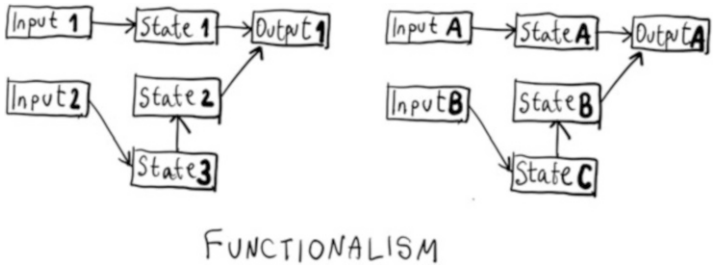

Fig. 8.2 Functionalism

by physics. From the perspective of physics, a "function" is nothing but a set of token-sequences. Consider three token-sequences: A′, B′, and C′; and suppose that A′ and B′ are of the same "functional kind" but C′ isn't. What fact makes this the case? If token-sequences are physical, then there are two options. Either (I) sequences A′ and B′ share a physical aspect (that C′ doesn't share with them); this is a reductive type-identity understanding of this case, analogous to the reductive type physicalist understanding of kinds of token-states, discussed above; or (II) A′ and B′ are heterogeneous, in which case in order to subsume them under the same "functional kind" we need a Sequences-Device, analogous to the one used above in the case of token-states. So Functionalism (like any approach that allows for multiple realizability) is either a version of reductive type-identity physicalism described by Flat Physicalism, or a version of token-dualism.

We conclude with Kim (2012, pp. 177–8):

> Token physicalism . . . is no physicalism – unless of course one lets it collapse to type physicalism.

8.7 The Autonomy of the Special Sciences

Fodor (1974) (and in a different way also Davidson, 1970[24]) wanted to guarantee genuine autonomy of the laws of the special sciences or some sort of freedom of the special sciences from physics. We have shown that the full-fledged reductive physicalist identity theory of Flat Physicalism is the only theory compatible with physicalism, according to which every token occurrence in the world is completely physical. If the special sciences' kinds aren't strictly identical to physical kinds (however complex), as in functionalism of all versions, then there is something *non-physical* in every *individual* token, so that *token-dualism* follows. Fodor (1974,

[24] Davidson's anomalous monism approach also implies straightforward token-dualism of the substance dualism form because one can show that it implicitly presupposes an outside *descriptor over and above* its ontological network of events that as it were creates the kinds: see our 2021b).

1997), wishing to accommodate multiple-realizability and ensure autonomy for the special sciences, opted for what he thought is token-physicalism, but which turns out to be *token-dualism*. Therefore, his route turns out to be *incompatible* with physicalism. As we shall briefly explain now, *Flat Physicalism, despite its reductive nature, surprisingly guarantees some particular sort of genuine autonomy for the special sciences*, which is in line with actual scientific practice.

As we said above, it is a consequence of statistical mechanics in the Flat Physicalism version that *every token* (state or sequence) has infinitely many aspects (i.e., macrovariables), but only some of them appear in our experience and in our theories. Again: we stress that in Flat Physicalism all the special sciences' kinds are macrovariables, namely partial descriptions of the actual microstate of the universe. This means that all of them are part and parcel of the microscopic blue-print of the universe "living" as it were at the fundamental level of reality which is all there is to the universe according to contemporary physics (therefore: there are no levels of reality; see the next section).

The proverbial Laplacian Demon, who knows the complete micro-trajectory of the universe and – if told which degrees of freedom of the universe are "us" – can calculate which are the macrovariables of our environment to which we are sensitive, and which are the macrovariables that make up our "measuring devices", namely our sense organs and brain states. Moreover, Laplace's Demon can predict which macrovariables these are, as well as which macroscopic laws of nature and regularities that we experience. That is, the Demon can say how our special sciences look like and what the special sciences' kinds and laws are. These facts are not mysterious (contra Fodor) as far as Flat Physicalism is concerned.

Here it is of crucial importance to notice two points: (i) Since macrovariables are in general partial descriptions of the microstate of the universe, it is a mathematical consequence that the time evolution of the special sciences' macrovariables will not in general even resemble the time evolution of the microstate of the universe as described by fundamental physics (e.g., quantum mechanics or quantum field theory). Therefore, we should only expect that the special sciences laws will look different, even radically different from the laws of fundamental physics. (ii) What we just said above in point (i) is a mathematical consequence of statistical mechanics (classical and quantum) in the version of Flat Physicalism. Therefore, there is absolutely no contradiction whatsoever between the idea that the special sciences laws may take forms that are *radically different* from the laws of fundamental physics and the fact that the special sciences are *fully reducible* to fundamental physics. In general, the time evolution (i.e., the laws) of the special sciences depend on *both* the time evolution of the full microstate of the universe and the partition of the state space into the macrostates (corresponding to the macrovariables).

In this sense one may say that the laws of the special sciences express certain *harmonies* between the microscopic full dynamics and the partition of states. Given the full micro-dynamics these harmonies may well be deterministic, stochastic (with probabilities that are different from the probabilities that appear in the micro-dynamics (e.g., in certain versions of quantum mechanics) or in the statistical mechanical account of the thermodynamics (e.g., the probabilities appearing in the

statistical mechanical account of the Second Law of thermodynamics). Moreover, it could even be that the harmonies pertaining to certain macrovariables that may be relevant for, e.g., brain science, cognitive sciences, psychology or other social sciences, will be completely *anomalous* in the sense of Davidson (1970), namely that they cannot be described by strict or even probabilistic laws of any sort at all! We give a more detailed analysis in our (2021a) of the way in which the harmony between the full micro-dynamics and the partition of states pans. The fact that these harmonies depend on the partition of states and not only on the micro-dynamics is the reason why despite the full reductive nature of this picture, the laws of special sciences like biology or psychology may look completely independent of the micro-dynamics and the macrovariables that appear in fundamental physics or in the special physical sciences, such as thermodynamics or even chemistry. Of course, it may be that the thermodynamic macrovariables do play a role also in biology or brain sciences (for example), but whether or not this is the case and whether or not macrovariables that are unfamiliar hitherto play a role in these sciences is a contingent matter that should be investigated empirically.

Here are some examples of special sciences' laws that may seem to resist reduction to fundamental physics, but that in fact are nothing but *harmonies* between the *micro-dynamics* and the *partition* to states.

1. **The Second Law of thermodynamics** according to which the entropy of an isolated system cannot decrease in time is temporally directed, despite the time symmetry of *all* the laws of fundamental physics.[25] This directedness can be accounted for in terms of non-temporal local asymmetries (see Hemmo & Shenker, 2019a).

2. **The laws of statistical mechanics** are probabilistic, and despite the deterministic nature of the laws of classical physics, the statistical mechanical probabilities describe objective features of the world (see e.g., Frigg & Hoefer, 2015; List and Pivato; Hemmo & Shenker, 2012, 2016).[26] If the micro-dynamics is taken to be quantum mechanical and stochastic (as in some versions of quantum mechanics), still the statistical mechanical probabilities are very different from the quantum mechanical ones because of the dependence on the partition of states to the thermodynamic macrovariables. In both cases, the probabilities in statistical mechanics can be accounted for in terms of the harmony between the micro-

[25] The Charge-Parity-Time (CPT) theorem and the violations of time-symmetry and charge-parity symmetries in the quantum electroweak theory are considered irrelevant for the workings of the brain, because of the high level of energy at which these violations occur. But, even if they are relevant, it is conjectured (see Atkinson, 2006) that the origin of these violations (from which the CPT theorem follows) is in the low entropy past hypothesis introduced in (quantum and classical) statistical mechanics. If this conjecture is true, the fact that the CPT theorem originates in the low entropy past is another example for what we call the autonomy of the special sciences!

[26] On some versions of quantum mechanics, the fundamental laws are probabilistic. But the quantum probabilities are different from the probabilities that appear in statistical mechanics; see (Hemmo & Shenker, 2012, Appendix).

trajectories and the partition of the state space into sets (the thermodynamic macrostates) according to the macrovariables to which the measuring devices are sensitive in the sense explained above (see Hemmo & Shenker, 2012, 2016 for an extensive discussion of macrostates and probability in statistical mechanics; and 2015a, b, 2019a, b, c, 2021a for the reductive account of special sciences laws on the basis of statistical mechanics).

3. **Classical physical computation:** In broad outline, a classical physical computation is a microphysical process in which a system is prepared at the initial time t0 in a state in which it has a certain property (macrovariable); it then undergoes a certain microphysical process according to its (specially built) parameters and constraints, that is, it satisfies the classical equation of motion F = ma; and then, at the final time *t*1, a certain property (macrovariable) is measured on it. The initial macrovariable is the input, and the final macrovariable is the output of the computation. For example, the initial macrovariable is prepared by pressing your keyboard at *t*0, and the final macrovariable is the state of your screen at *t*1. This means that a classical physical computation is an implementation of a certain harmony between the microdynamics and the partition of states of the computing system (sometimes called value assignment) induced by the observer (or user). This harmony results in macroscopic dynamics (or transitions between macrovariables) that may be wildly different from the microdynamics, exactly as in the two examples above. More generally, we have shown elsewhere that this kind of harmony between microdynamics and partition to macrostates is the physical underpinning of Putnam's (1988, p. 121) theorem according to which: "Every ordinary open system is a realization of every abstract finite automaton."[27] This implies two things that are highly relevant to our discussion here: First: Putnam's theorem is in fact a consequence or a theorem of statistical mechanics. Second: since every classical microstate has *infinitely* many *macrovariables* that give rise to infinitely many partitionings of the state space, corresponding to infinitely many value assignments, every computing system in fact simultaneously implements infinitely many computations. We cannot go here into the details of these results. Our point in this example is to argue that classical physical computation is a special science with autonomy of exactly the kind we spelled out above and this autonomy is perfectly consistent with the full reduction of classical computation to the microdynamics and partition of states, as envisaged by Flat Physicalism. In fact, this autonomy is a consequence of the reduction to physics.

4. **Quantum computation:** A quantum computation is the following counterpart of the classical one.[28] A system undergoes an evolution which is, in non-

[27] Requiring certain physical input and output adds some constraints, but even in that case Putnam's claim is quite strong.

[28] On the main ideas of quantum computations see e.g., Feynman, 1982, 1996; Nielsen & Chuang, 2010, pp. 171–215, Mermin, 2007, Pitowsky, 1990).

> relativistic quantum mechanics, the Schroedinger equation. The details of this equation are fixed by the system's parameters and constraints. At time t0 the system is prepared in the initial quantum state, which is the computation's input; the system then evolves from time t0 to t1. At time t1 a certain quantum mechanical observable is measured on it, and this measurement's outcome is the computation's output. For example, if the input and output are described in the same basis, then the Schroedinger evolution induces a change in the amplitudes (typically, an amplification of one of the amplitudes, that is designed to be the computation's outcome with high probability). The great achievement of the discovery of algorithms for quantum computing is finding the 'right' Schroedinger evolution and the 'right' preparation and measurement that will allow implementing certain desired computations. The difference between classical and quantum computers lies wholly in the different physics, and as is well known, due to this difference quantum computers can implement certain computations much faster than classical ones. (For an overview of quantum computation, philosophical outlook, and references, see e.g., (Hagar, 2003, 2007; Hagar & Cuffaro, 2019; Cuffaro, 2012).

In these rough terms, the efforts in the contemporary research on quantum computation are, first, to write such quantum algorithms, i.e., to find the right Schroedinger equation and input and output observables that will implement a desired computation; and second, to find physical systems on which this quantum dynamics can be implemented. In the latter, a great challenge faced nowadays is to design systems that will undergo these evolutions without interruptions, mainly by decoherence interactions with an external environment that suppress the interference terms in the superposition in the decoherence basis (and therefore also suppress effectively the entanglement between different degrees of freedom). As a result of decoherence, any computation that is carried out by macrovariables that commute with the decoherence Hamiltonian becomes effectively classical, so that the efficiency of the quantum contribution to the computation is reduced almost completely (for these macrovariables). Although there are some successful results in this context, these tasks turn out to be quite challenging, and enormous efforts and resources are nowadays being invested in attempts to build a quantum computer with some significant number of qubits.

One last point about this example: Putnam's theorem, which as we said, is a theorem of classical statistical mechanics, is carried over (mutatis mutandi) to the quantum case. But in quantum mechanics it turns out that the multiplicity of computations is much more radical than in the classical case. First: It is a mathematical fact that the quantum state of the computing system has many macrovariables that, given the right partitions, turn out to implement many computations (this is the quantum analogue of the classical multiple-computations theorem). Second and moreover: it is a mathematical fact that the quantum state and the Schroedinger evolution of it can be equally described in infinitely many bases of Hilbert space corresponding to the infinitely many (micro-) observables one may wish to measure. (For example, if in one basis the input amplitudes are uniform and the output has one of them amplified,

in another basis this may be different.)[29] This is why we stressed above that decoherence interactions have a disruptive effect on the quantum computation only relative to observables that commute with the decoherence basis. That is, even if the computing system undergoes environmental decoherence, one may be able to carry out a genuinely efficient quantum computation by finding the right observables, i.e., certain observables that do not commute with the decoherence Hamiltonian, and therefore will be sensitive to the interference terms in the corresponding bases of Hilbert space. These bases are no less real than the decoherence basis at any time during the Schroedinger evolution of the quantum state of the system because of the basis symmetry of Hilbert space.[30]

To sum up this issue of autonomy of the special sciences: As a matter of principle, we are *unable* to discover the facts about the harmony between the microdynamics and partition of states by doing fundamental physics. The reason is not only because of the immense complexity of the world, but also because we ourselves are part of it, and the very act of making predictions is a *physical* process in our brain, that is, in the world that is to be predicted. This process is presumably *macroscopic* involving macrobariables of our brain, which is the subject matter not of fundamental physics, but rather of special sciences such as brain science, cognitive science and psychology. Thus, the only way in which we can learn which are the aspects (macrovariables) of the world to which we are sensitive and what are their regularities, as well as which are the macrovariables of our brain and body which are in fact correlated with these outside macrovariables, is by doing special sciences, not fundamental physics![31] We cannot discover these macrovariables by only looking at our microphysic physical theory, however fundamental. Physics cannot do this. This is the fact that underlies the genuine autonomy of the special sciences. Unlike the mysterious nature of this autonomy as described by, for example, Fodor (1974, 1997), according to Flat Physicalism the autonomy of the special sciences is a fact that has a straightforward physical explanation. The fact that the special sciences' laws are autonomous in the above sense supports our view that a psycho-physical *identity* theory doesn't entail eliminativism with respect to the special sciences' kinds.

[29] The multiplicity here is due to the fact that the quantum state is invariant under basis transformation. Given the input and output states, relative to each basis expanding the quantum state at the intermediate times, the system may be said to implement a *different* computation. This is to be distinguished from the idea that *given the computational basis*, the quantum mechanical process may be said to consist of 'parallel' computations carried out along the branches of the state (in the computational basis; see discussion and criticism in Cuffaro, 2012).

[30] We argue elsewhere (see our 2020, 2021d) that this is why decoherence by itself does not solve the so-called 'preferred basis problem' in the many worlds interpretation of quantum mechanics, unless one adds facts or laws over and above the Hilert space structure that give preference to the decoherence basis.

[31] Likewise: the only way in which we can discover other aspects or macrovariables that are accessible only via complex measuring devices and that may satisfy interesting and useful regularities, is the job of the special sciences.

8.8 Levels of Reality

According to Flat Physicalism (as its name indicates) the world is absolutely, categorically, unequivocally, *flat*: there are no so-called "*levels of reality*" ("levels of explanation" is only a *façon de parler* and has no bearing on reality). In this paper we have assumed that: (i) reality is as described by physics, the toy model being a world fully and correctly described by classical mechanics; and that (ii) this is all that there is in reality, that is, the description of reality as given by microphysics is complete. Physical aspects of the microstate of the universe are parts of this complete reality, they don't stand for additional structure or facts over and above the complete physical microstructure. Sets (in the state space postulated by physics) of microstates that share aspects (called macrostates) don't exist in the world, for according to our best theories of contemporary physics, each such set consists of one *actual* microstate and (infinitely) many *counterfactual* ones, which aren't in the world.

The denial of levels in Flat Physicalism is quite radical. The prevalent sentiment in contemporary philosophical literature on this matter is represented in the following quotation (brought in a different context – that of asking whether the relation of "grounding" is fundamental).

> I have no knockdown argument against the claim that the world is flat. But every fiber of my being cries out in protest. ... The true flatworlder ... denies that there are *any* non-fundamental properties, and, indeed, ... she denies that there are states of affairs, she denies that there are sets, she denies that there are people. ... any version of flatworldism will be radically revisionary. I repeat that I have no real argument against it. I will simply say that flatworldism is, to borrow a colorful word from a friend, "crazypants". An imprecise complaint to be sure, but it is my complaint nonetheless. It is a cousin of the incredulous stare. (Bennett, 2011, p. 28).

Bennett (2011) is very clear that her rejection of "flatworldism" is (at this preliminary stage, at least) a matter of intuition (unlike, for example, the argument in Fodor, 1974, 1997, mentioned above). Our intuition, that guides us in this paper, is the opposite one: we find it hard to see how levels of reality can come to exist, and what sort of existence is meant by existence in different levels. The very fact that opposing intuitions are available makes it easier to discuss the matter. Here we shall not argue for (nor against) flat reality: our main task is to show what a flat reality amounts to and how it is a consequence of Flat Physicalism as described above.

By "*levels of reality*" we mean (in this paper) this: A multi-level structure of reality is one in which there are matters of fact at a relatively high level that aren't part of a relatively low level. Not vice versa: a case in which there are matters of fact at a low level that aren't part of a higher level will not be treated, in this paper, as one in which reality is multi-leveled; below we explain why this is so.

Terminological remarks: (1) In the literature there are various other notions denoted by the term 'levels'; for example, there are 'levels of explanation', and other notions. We don't address them (see, for example, Craver, 2007; Craver & Bechtel, 2007; Frigg & Hoefer, 2015, List & Pivato, 2015; Bechtel, 2016; List, 2019). (2) The terms '*matters of fact*' or '*facts*' are intended (in this paper) to be wide and

inclusive, and the way we think of them will hopefully become clear as we proceed. (3) The term '*relatively*' is intended to put to the side, in this paper, the question of whether or not there is a fundamental level of reality. Hereafter we shall omit the term 'relatively' in this context, for simplicity. To make things even simpler we shall assume for the sake of the argument that there is a fundamental level, and call that level "physical".

Often when people discuss the multi-levels structure of reality (either in the above sense or in other senses) they do so in the context of characterizing reality in terms of two concepts: whether or not *multiple-realizability* of special science kinds by physical kinds obtains, and whether or not *supervenience* of special science kinds on physical kinds obtains.

Consider Fig. 8.3: in the two cases described there supervenience obtains, but whereas in case 2 multiple-realization is allowed or possible (the special kind X is realized by the physical kinds A and B), in case 1 it isn't. Supervenience is often taken to be the hallmark of physicalism, giving physics a preferred status over all the special sciences (see for example, Kim, 2012). In terms of Flat Physicalism, in case 1, tokens 1 and 2 are both of the kind X in virtue of their sharing a physical fact, specifically: they share the aspect A of their microstate. They are partially identical, and the aspect in which they are identical *is* the kind X. Tokens 3 and 4 don't have this aspect, and it is therefore that they don't belong to the kind X. To make our point clear, here is an example.

(Terminological remark: Some people call case 1 "multiple-realization" by which they mean that the property X is shared by the different tokens 1 and 2 that, although they share the aspect A, differ in other aspects. There is no point in arguing about terminology, as long as the core of the matter is understood, confusion between cases 1 and 2 is avoided, and the way in which genuine multiple realization is a form of dualism is understood. Since *we* find this terminology confusing, we avoid it, and emphasize that 1 is a case of strict identity (between the kind X and the aspect

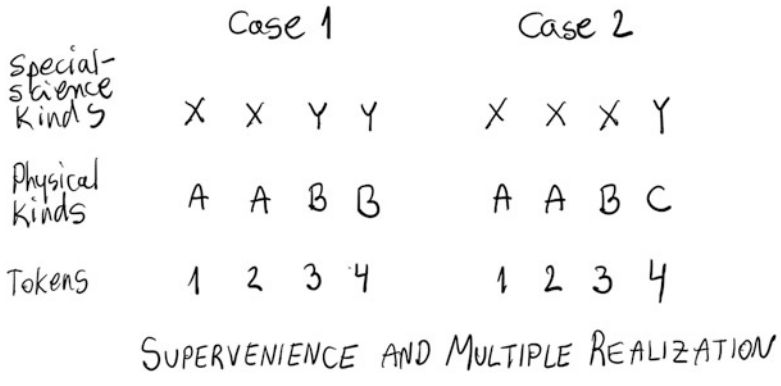

Fig. 8.3 Supervenience and multiple realization

A that is present in each of the tokens 1 and 2 but not 3 and 4), whereas Case 2 is a case of genuine multiple-realization.)

Case 1 is the *only* case in which reality can be flat, that is, without multiple levels (in the above sense of this term).[32] In case 2, by contrast, reality is leveled, since in order to make tokens 1, 2, and 3 belong to the same kind X despite the fact that token 3 doesn't share a physical aspect with tokens 1 and 2, another fact – beyond that of the complete physical tokens – needs to be brought in. This additional fact might be The Device mentioned above; but if The Device is us, and its pointer states are our brain-cum-mental states, then the extra non-physical fact has to be external to physics. Adding a non-physical fact brings about another layer, so to speak, of reality.

8.9 Conclusion: Functionalism as a form of Dualism, and Some Consequences

Are there facts in the world over and above those described by physics? This is a question of fact, and we don't know what is the truth about the world; hence we don't attempt to prove reductive physicalism nor to disprove functionalism or any other form of dualism. Our task is, rather, *to identify which views are forms of dualism and which aren't*, and in particular, to point out that views that call themselves "non-reductive physicalism" aren't physicalist *at all*, since they posit the existence, as part of each individual token, of certain non-physical facts, that fix the special-sciences sets to which this token belongs.

The conclusion from what we have seen so far is that if multiple-realization is allowed, *even if supervenience is required*, then the matters of fact that make it the case that a certain tokens (but not others) belong to certain kinds (rather than others) aren't, and cannot be, physical matters of fact, but must be non-physical matters of fact. Importantly, in the case in which both multiple-realization and supervenience hold, the non-physical matters of fact must be *part of every token*, so that it would be wrong to say that each token is physical and nevertheless the (so-called high-level) kinds cannot be fully accounted for by the physics of the token. We called this *token-dualism*. And so, our claim is that *allowing for multiple-realization, even with supervenience, entails (or just is) token-dualism*.

The term *dualism* isn't meant pejoratively; for all we know dualism might be true about the world! It is very rich and has a variety of meanings in the literature. Our aim is to show that some approaches which present themselves as forms of physicalism are, in fact, forms of dualism. This may lead to conceptual confusion. And so, *our task is a clarificatory one: to clarify the metaphysical commitments of some contemporary approaches*: which views are dualistic and which aren't in the following sense: In this paper we understand *dualism* along the lines of the slogan

[32] See also arguments for a flat picture in (Bechtel, 2016).

often used to describe *physicalism*: if physicalism is the view that *everything is physical*, then dualism is the view that *some things aren't physical*.

While we did not focus in this paper on direct arguments supporting reductive physicalism, it may be significant to add a word of motivation here (but our argument is totally independent of this motivation). The view that accepts, seriously and authentically, the idea of multiple-realizability, is a hindrance to the advancement of science, because it says that there is no point in searching for the common physical feature of cases that appear to be of multiple-realization. Fortunately, working scientists don't accept this idea and continue with their research, and whenever they find shared feature they see it as progress – a judgement that would be meaningless for friends of multiple-realizability (see for example, cases described in Polger & Shapiro, 2016).

Our main target, in addressing forms of non-reductive physicalism, is the various forms of functionalism, including computational functionalism. Since functionalism requires, or is at least compatible with, multiple-realizability of functional kinds (for example, causal kinds) by physical kinds (including biological and psychological kinds), our proof shows that *functionalism is a form of token-dualism*, in the above sense of the terms. This argument only depends on multiple-realizability. It is independent of other features of functionalism, in particular computational functionalism, which are also in tension with physicalism, such as the multiple-computations thesis (also originally due to Putnam, 1988). According to this well-established thesis (even only in its weakest form[33]), the time evolution of every open macroscopic system implements more than a single computation. This result means that functionalism implies a violation of *supervenience*. There is a debate about whether or not this kind of failure of supervenience is problematic (see, for example, Chalmers, 2012, Shagrir, 2012); many contemporary thinkers in this field take this result as seriously challenging the functionalist theory of mind. It has been recently shown (see Hemmo & Shenker, 2019b) that the only solution to this problem that is compatible with physicalism is a *type-identity* theory in which the question of which function (or computation) is implemented by a given system is *uniquely* determined by the sequence of brain states which are identical to the mental states of an observer. But our argument here that functionalism implies token-dualism doesn't depend on any considerations related to multiple-functions or multiple-computations, but rather only on the assumption that mental kinds are multiply-realizable by heterogeneous physical kinds.

We have shown that this assumption alone implies that in each and every token-sequence of microstates realizing a high-level functional kind (or implementing a certain computation), there must be some non-physical sequence of facts in virtue of which it belongs to that functional kind. This conclusion pertains to *all* forms of functionalism, including computational functionalism. Putnam (1975) famously

[33] This result holds even if one adds various counterfactual and other constraints on what counts as a physical implementation of a function; see, for example, (Godfrey-Smith, 2009), (Schuetz, 2012), (Piccinini, 2015), and (Piccinini, 2017) for a recent overview of this issue.

said that "the functional-state hypothesis is *not incompatible with* dualism!".[34] We have argued in this paper for a much stronger conclusion: the "functional-state hypothesis" *entails* token-dualism or *is a form of* token-dualism, and is therefore *incompatible with physicalism*.

Acknowledgement This research is supported by the Israel Science Foundation, grant number 1148/2018 and grant number 690/2021. For valuable comments and criticism we thank all the participants of the workshop on the Multi-Level Structure of Reality held at the Israel Institute for Advanced Studies at the Hebrew University and the University of Haifa on 26–30 May, 2019.

References

Atkinson, D. (2006). Does quantum electrodynamics have an arrow of time? *Studies in History and Philosophy of Modern Physics, 37*(3), 528–554.
Bechtel, W. (2016). Explicating top-down causation using networks and dynamics. *Philosophy of Science, 84*, 253–274.
Ben-Menahem, Y. (2018). *Causation in science*. Press.
Bennett, K. (2011). By our bootstraps. *Philosophical Perspectives, 25*, 27–41.
Bickle, J. (2010). Has the last decade of challenges to the multiple realization argument given aid and comfort to psychoneural reductionists. *Synthese, 177*, 247–260.
Bickle, J. (2013). Multiple realizability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019 ed.). URL = https://plato.stanford.edu/archives/spr2019/entries/multiple-realizability/
Brown, R. G., & Ladyman, J. (2019). *Materialism: A philosophical inquiry*. Routledge.
Chalmers, D. (1996). *The conscious mind*. Press.
Chalmers, D. (2012). The varieties of computation: A reply. *Journal of Cognitive Science, 13*, 211–248.
Chang, H. (2012). *Is water H2O?* Springer.
Crane, T. (2017). How we can be. *Times Literary Supplement, 5956*, 7–8.
Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.
Craver, C., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*, 547–563.
Cuffaro, M., & E. (2012). Many worlds, the cluster-state quantum computer, and the problem of the preferred basis. *Studies in History and Philosophy of Modern Physics, 43*, 35–42.
Davidson, D. (1970). Mental events. In D. Davidson (Ed.) (1980), *Essays on Actions and Events* (pp. 207–227). University of California Press.
Elpidrou, A. (2018). The character of physicalism. *Topoi, 37*, 435–455.
Feynman, R. (1982). Simulating physics with computers. *International Journal of Theoretical Physics, 21*(6/7), 467–488.
Feynman, R. (1996). In J. G. Hey & W. Allen (Eds.), *Feynman lectures on computation*. Addison-Wesley.
Frisch, M. (2015). *Causal reasoning in physics*. Cambridge University Press.

[34] The reason for this is, in his words, that: "Although it goes without saying that the hypothesis is "mechanistic" in its inspiration, it is a slightly remarkable fact that a system consisting of a body and a "soul," if such things there be, can perfectly well be a Probabilistic Automaton."

Firt, E., Hemmo, M., & Shenker, O (2021). Hempel's Dilemma: Not only for physicalism. *International Studies in Philosophy of Science*, under review R&R.

Frigg, R. (2008). A field guide to recent work on the foundations of statistical mechanics. In D. Rickles (Ed.), *The Ashgate companion to contemporary philosophy of physics* (pp. 99–196). Ashgate.

Frigg, R., & Hoefer, C. (2015). The best humean system for statistical mechanics. *Erkenntnis, 80*, 551–574.

Fodor, J. (1974). Special sciences: Or the disunity of science as a working hypothesis. *Synthese, 28*, 97–115.

Fodor, J. (1997). Special sciences: Still autonomous after all these years. *Nous, 31*, 149–163.

Godfrey-Smith, P. (2009). Triviality arguments against functionalism. *Philosophical Studies, 145*, 273–295.

Hagar, A. (2003). A philosopher looks at quantum information theory. *Philosophy of Science, 70*, 752–775.

Hagar, A. (2007). Quantum algorithms: Philosophical lessons. *Minds & Machines, 17*, 233–247.

Hagar, A., & Cuffaro, M. (2019). Quantum computing. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.).

Hemmo, M., & Shenker, O. (2012). *The road to Maxwell's Demon: Conceptual foundations of statistical mechanics*. Cambridge University Press.

Hemmo, M., & Shenker, O. (2013). Entropy and computation: The Landauer-Bennett thesis reexamined. *Entropy, 15*, 3387e3401.

Hemmo, M., & Shenker, O. (2015a). Probability and typicality in deterministic physics. *Erkenntnis, 80*, 575–586.

Hemmo, M., & Shenker, O. (2015b). The emergence of macroscopic regularity. *Mind & Society, 14*(2), 221–244. https://doi.org/10.1007/s11299-015-0176-x

Hemmo, M., & Shenker, O. (2016). *Maxwell's Demon*. Oxford Handbooks Online.

Hemmo, M., & Shenker, O. (2019a). The past hypothesis and the psychological arrow of time. *British Journal for the Philosophy of Science*, axz038. https://doi.org/10.1093/bjps/axz038.

Hemmo, M., & Shenker, O. (2019b). Two kinds of high-level probability. *The Monist, 102*, 458–477.

Hemmo, M., & Shenker, O. (2019c). The physics of implementing logic: Landauer's principle and the multiple-computations theorem. *Studies in History and Philosophy of Modern Physics* , . Forthcoming. https://doi.org/10.1016/j.shpsb.2019.07.001

Hemmo, M., & Shenker, O. (2020). Why the Many Worlds interpretation of quantum mechanics needs more than Hilbert space structure. In R. Peels, J. de Ridder, & R. van Woudenberg (Eds.), *Scientific challenges to common sense philosophy* (pp. 61–70). Routledge, Taylor & Francis.

Hemmo, M., & Shenker, O. (2021a). Flat physicalism. *Theoria*. Forthcoming.

Hemmo, M., & Shenker, O. (2021b). A dilemma for Davidson's anomalous monism. In Y. Ben-Menahem (Ed.), *Laws of nature*. Springer. Forthcoming.

Hemmo, M., Shenker, O. (2021c) A challenge to the second law of thermodynamics from cognitive science and vice versa. *Synthese* (.), 1–31. https://doi.org/10.1007/s11229-020-03008-0.

Hemmo, M., & Shenker, O. (2021d). Why decoherence does not solve the preferred basis problem in the many-worlds interpretation of quantum mechanics. *Synthese* (R&R, second revision).

Hoefer, C., & Marti, G. (2019). Water has a microstructural essence after all. *European Journal for Philosophy of Science, 9*(12). https://doi.org/10.1007/s13194-018-0236-2

Kim, J. (2012). The very idea of token physicalism. In G. Simone & H. Christopher (Eds.), *New perspectives on type-identity* (pp. 167–185). Cambridge University Press.

Kripke, S. (1980). *Naming and necessity*. Wiley.

Ladyman, J., & Ross, D. (2007). *Every thing must go*. Oxford University Press (with D. Spurrett and J. Collier).

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

List, C., & Pivato, M. (2015). Emergent chance. *Philosophical Review, 124*(1), 119–152.

Mermin, D., & N. (2007). *Quantum computer science: An introduction*. Cambridge University Press.

Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information*. Cambridge University Press.
Papineau, D. (1993). Physicalism, consciousness and the antipathetic fallacy. *Australasian Journal of Philosophy, 71*, 169–183.
Piccinini, G. (2015). *Physical computation*. Press.
Piccinini, G. (2017). Computation in physical systems. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2017 ed.). URL = https://plato.stanford.edu/archives/sum2017/entries/computation-physicalsystems/
Pitowsky, I. (1990). The physical church thesis and physical computational complexity. *Iyyun: The Jerusalem Philosophical Quarterly, 39*, 81–99.
Polger, T., & Shapiro, L. (2016). *The multiple-realization book*. Oxford University Press.
Portides, D. (2019). Idealizations and abstraction in scientific modelling. *Synthese*. https://doi.org/10.1007/s11229-018-01919-7
Psillos, S. (2009). Knowing the structure of nature. *Macmillan.*
Putnam, H. (1975). The nature of mental states. In H. Putnam (Ed.), *Mind, language and reality* (pp. 429–440). Cambridge University Press (1975). Originally published as "Psychological Predicates", in: William H. Capitan and Daniel D. Merrill (Eds.), Art, Mind and Religion, pp. 37–48, Pittsburgh, PA: University of Pittsburgh Press (1967).
Putnam, H. (1988). *Representation and Reality*. MIT Press.
Ramsey, W. (2019). Eliminative materialism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2019 ed.). URL = https://plato.stanford.edu/archives/spr2019/entries/materialism-eliminative/
Russell, B. (1925). Preface. In F. Lange (Ed.), *The history of materialism and criticism of its present importance* (E. C. Thomas, Trans). Routledge.
Schuetz, M. (2012). What is it not to implement a computation: A critical analysis of Chalmers' notion of implementation. *Journal of Cognitive Science, 13*, 75–106.
Shenker, O. (2017a). Foundations of statistical mechanics: Mechanics by itself. *Philosophy Compass*. https://doi.org/10.1111/phc3.12465
Shenker, O. (2017b). Foundations of statistical mechanics: The auxiliary hypotheses. *Philosophy Compass*. https://doi.org/10.1111/phc3.12464
Shenker, O. (2018). Foundations of quantum statistical mechanics. In E. Knox & A. Wilson (Eds.), *Routledge companion to the philosophy of physics*. Routledge. Forthcoming, expected in 2020.
Shagrir, O. (2005). The rise and fall of computational functionalism. In Y. Ben-Menahem (Ed.), *Hilary Putnam: Contemporary philosophy in focus*. Cambridge University Press.
Shagrir, O. (2012). Can a brain possess two minds? *Journal of Cognitive Science, 13*, 145–165.
Sklar, L. (1993). *Physics and chance*. Cambridge University Press.
Smart, J. J. C. (1959). Sensations and brain processes. *The Philosophical Review, 68*, 141–156.
Smart, J. J. C. (2017). The mind/brain identity theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2017 ed.). URL: https://plato.stanford.edu/archives/spr2017/entries/mind-identity/
Tiehen Justin. (2018). Physicalism. *Analysis, 78*, 537–551.
Uffink, J. (2007). Compendium to the foundations of classical statistical physics. In J. Butterfield & J. Earman (Eds.), *Handbook for the philosophy of physics*, Part B, pp. 923–1074.
Wallace, D. (2001). *Implications of quantum theory in the foundations of statistical mechanics*. Unpublished manuscript. Preprint available at http://philsci-archive.pitt.edu/410/

Chapter 9
Levels and Mechanisms: Reconsidering Multi-level Mechanistic Explanation

Stavros Ioannidis and Stathis Psillos

Abstract In this paper we present and defend a causal account of multi-level mechanistic explanation by reconsidering the relationship between levels and mechanisms. We argue for three main claims: (i) that biological mechanisms are causal pathways, (ii) that levels and mechanisms are distinct notions and (iii) that levels of nature and of multi-level explanations are levels of composition. According to our view, multi-level mechanistic explanations identify causal pathways that contain components from multiple levels of composition. We contrast our view with Craver's well-known account of multi-level explanation and his notion of levels of mechanisms. We discuss various biological examples of multi-level explanation in order to motivate and illustrate our view, and argue that, in contrast to a view such as Craver and Bechtel's, it allows for interlevel causation.

9.1 Introduction

During the last few decades mechanistic approaches have witnessed great development within philosophy of science. According to New Mechanism, as the current version of the mechanistic philosophy has been called, mechanisms are the key category (both ontologically and methodologically) for understanding the nature of the life sciences. Such central issues as scientific explanation, scientific discovery, the relationships among scientific fields, as well as the metaphysics of science, are being reconsidered in mechanistic terms (Craver & Tabery, 2015).

S. Ioannidis (✉)
Department of History and Philosophy of Science, National & Kapodistrian University of Athens, Athens, Greece
e-mail: sioannidis@phs.uoa.gr

S. Psillos
Department of History and Philosophy of Science, National & Kapodistrian University of Athens, Athens, Greece
e-mail: psillos@phs.uoa.gr

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_9

According to a widespread view associated with New Mechanism, levels of nature typically invoked in explanations in life sciences are levels of mechanisms. On Carl Craver's (2007) popular account, the relation between mechanistic levels is viewed in terms of the relation between the mechanism as a whole and its components; in turn, this relation is considered to be a non-causal dependency relation, to be viewed in terms of mutual manipulability. So, Craver's mutual manipulability account serves both to give an account of the non-causal relations between the components and the whole mechanism, and to ground a hierarchy of mechanistic levels.

The aim of this paper is to reconsider the relationship between mechanisms and levels, and present a new account of multi-level mechanistic explanation by examining various examples of multi-level mechanisms. The main ingredients of this alternative view are three basic claims: (i) biological mechanisms are causal pathways, (ii) levels and mechanisms are distinct notions and (iii) levels of nature and of multi-level explanations are levels of composition. Our key idea is that whatever contributes to the production of a phenomenon is part of the same causal pathway; but causal pathways may contain entities that belong to multiple levels of composition. While multi-level explanations refer to entities from various compositional levels, compositional levels as such do not matter for explanation; what matters is the existence of a causal pathway. This view of multi-level explanation is in stark contrast to the view developed in Craver (2007) and in Craver & Bechtel (2007), according to which 'levels' in multi-level explanations are levels of mechanisms and mechanistic multi-level explanations are instances of constitutive explanations.

We begin (Sect. 9.2) by discussing the notion of levels of composition and argue that this is a typical way to understand levels talk in biology. We then (Sect. 9.3) contrast levels of composition with Craver and Bechtel's levels of mechanisms and discuss two main problems of the levels of mechanisms account, i.e. that it presupposes a conception of mechanism that is not very useful in biological practice (the 'constitutive' conception) and that it leads to a very narrow notion of levels. In Sect. 9.4 we present our account of mechanism as a concept-in-use in biology, which we call *Causal Mechanism*. Its core idea is that mechanisms in biology are causal pathways. We then (Sect. 9.5) present our alternative view of multi-level mechanistic explanation, which we illustrate by various examples (Sect. 9.6): the mechanisms of apoptosis, scurvy and type 2 diabetes. In the last section (Sect. 9.7) we argue that our view allows for interlevel causation.

9.2 Levels of Composition in Biology

In science and philosophy one can find various different notions of levels (see Craver, 2007, Eronen & Brooks, 2018). In life sciences, however, the notion of levels is very often used in the sense of levels of organisation or composition. The idea here is that biological objects form a hierarchy defined by part-whole

or compositional relationships; for example, molecules such as proteins and DNA compose cells, which compose tissues, which compose organs, that are parts of organisms, which are parts of ecosystems. Cells, on this view, are at a higher level than DNA and proteins, but at a lower level than tissues. Such a view of levels goes back to the account offered by Oppenheim and Putnam (1958) and takes it that the level of an entity (such as a cell) is a more or less objective matter, since relations between levels are grounded in mereological or compositional relations between things, which are objective features of the world.[1]

Levels of composition can be contrasted with what we will call levels of scope. Whereas levels of composition concern entities such as biological objects that form mereological hierarchies, levels of scope concern the laws of nature that govern these entities. Laws of physics, for example, govern particles, but also cells, organisms, etc.; while laws of biology and of other special sciences (or whatever plays the role of laws in these sciences) govern specific entities, e.g. cells and organisms.[2] Laws of physics, then, have a much wider scope than laws of biology; and hence we commonly say that physics itself has a much wider scope than other sciences. Physical laws govern entities at many (and perhaps all) levels of composition, whereas laws of biology and other special sciences (or whatever plays the role of laws in these sciences) govern entities at specific levels of composition. We have then one kind of hierarchy, formed by compositional relations; and another one, given by levels of scope.[3] The latter hierarchy can be used to arrange scientific disciplines hierarchically, according to the generality of their laws (i.e. according to the range of entities over which their laws range).

This is of course a sketchy picture; there are various issues that need to be addressed in a more fully developed account of levels. A first point to note is that the resulting hierarchy need not be a simple linear structure, as the one given by Oppenheim and Putnam. More complex hierarchies formed by compositional relations are possible, where the resulting hierarchy has a branching structure as that described by Wimsatt (1976), where a common atomic level gives rise to two main branches, one leading to planets, stars and cosmological objects, the other (very complex one) to organisms and societies. In such a structure, it will not be generally the case that for every two objects, either one will be at a higher/lower level than the other, or they will be both at the same level. For example, both planets and organisms are at a higher level than particles (since both are composed of particles), but it is not the case that planets are at a higher/lower level than organisms (since they are not composed of them/do not compose them), and they are not at the same level either. What this means, is that levels are not 'monolithic'

[1] We leave it as an open question whether there is a fundamental level.

[2] Even if, as new mechanists argue, mechanisms rather than laws are central in life sciences, this does not imply that there are no laws in biology (for accounts of biological laws see for example Waters (1998) and Mitchell (2000)).

[3] What levels there are may depend on what kind of regularities there are, as in Wimsatt's account of levels of organisation, where levels are taken to be 'local maxima of regularity and predictability' (1976, 209).

as in the Oppenheim and Putnam picture, but have a more 'local' character. Even in more complex hierarchical structures, however, which are much more realistic than Oppenheim and Putnam's simple picture, we have hierarchies of levels formed by mereological or compositional relations, with entities at different levels being governed by different kinds of laws (with some laws having a wider scope than others). Levels of composition and scope, then, do not depend on the existence of a linear hierarchy.

Moreover, we need not assume that the hierarchy of levels of composition has to be 'nested' in the sense that entities at a specific level will be composed only of entities at the next lower level. In biological hierarchies, in particular, this is not usually the case: for example, a cell is not composed only of cell organelles, but also of various kinds of molecules (see also Potochnik & McGill, 2012). Lastly, we need not assume that there will be a neat correspondence between scientific disciplines and levels, as again supposed by Oppenheim and Putnam. A given level of composition may be studied by various scientific disciplines, and one scientific discipline may correspond to various compositional levels (see also Craver, 2007). In general, we should not expect that the structure of levels will correspond to a similar structure of scientific disciplines.

Apart from how exactly a fully developed account of levels of composition will look like, we take it that a picture such as the one sketched above emerges from our best science. The scientific image contains entities of various sizes and degrees of complexity, that form hierarchies ranging from subatomic particles to superclusters of galaxies, in terms of size, and from subatomic particles to brain and organisms and other higher-level entities in terms of complexity. The existence of levels of composition (and of scope) is thus uncontroversial.

But moreover, and what mainly concerns us in this paper, levels of composition provide a way to make sense of levels talk in life sciences. Biologists often use expressions like 'levels of complexity' or 'levels of organisation' and talk about the 'molecular', the 'cellular', the 'organismic' and the 'ecological' level. Each such level is characterised by a distinctive set of entities and kinds of interactions. The genetic level, for example, involves genes, their expression and interactions among them, whereas biological parts like limbs and brains belong to a higher organisational level. Each level has its own principles of operation and higher levels cannot be 'reduced' to lower ones, but the levels are not independent: the behaviour of entities at higher levels are taken by biologists to depend on what happens at lower levels; moreover, higher levels can constrain what happens at lower levels.

Here is for example how developmental biologist Scott Gilbert characterises levels of organisation in his well-known textbook on developmental biology:

> The properties of a system at any given level of organization cannot be totally explained by those of levels 'below' it. Thus, temperature is not a property of an atom, but a property that 'emerges' from an aggregate of atoms. Similarly, voltage potential is a property of a biological membrane but not of any of its components. Higher-level properties result from lower-level activities, but they must be understood in the context of the whole (2010, 618).

Gilbert generalises this picture:

> Parts are organized into wholes, and these wholes are often components of larger wholes. Moreover, at each biological level there are *appropriate rules*, and one cannot necessarily 'reduce' all the properties of body tissues to atomic phenomena. ... When you have an entity as complex as the cell, the fact that quarks have certain spins is irrelevant. This is not to say, however, that each level is independent of those 'below' it. To the contrary, laws at one level may be almost deterministically dependent on those at lower levels; but they may also be dependent on levels 'above' (2010, 620, emphasis added).

Such a view of levels of organisation is a traditional one among biologists. Gilbert (2010, 620) quotes Joseph Needham (1943), who writes:

> The deadlock [between mechanism and vitalism] is overcome when it is realized that every level of organization has its own regularities and principles, not reducible to those appropriate to lower levels of organization, nor applicable to higher levels, but at the same time in no way inscrutable or immune from scientific analysis and comprehension.

Gilbert's levels of organisation are levels of composition, since a membrane is at a higher level than the molecules that compose it. Moreover, higher levels have their distinctive properties (e.g. temperature, voltage potential) that might not be possessed by their constituent lower-level entities; as a result, higher levels are governed by a distinctive set of laws or rules. At the same time, they depend on lower levels, but may influence them in turn. We see then that the concept of levels of composition is an important notion in biology. This leads to an a posteriori argument in favour of the existence of levels of composition as objective features of the world: according to our best science, the world (the biological world in particular) contains a hierarchy of levels of composition.[4]

9.3 Craver and Bechtel on Levels of Mechanisms

We have seen that the view that levels in biology are levels of composition is a traditional and very plausible one. However, according to Craver's (2007) very influential account, levels in biology (and in neuroscience in particular) are not simply levels of composition, but levels of mechanisms.

At the core of Craver's account lies his conception of a 'constitutive' mechanism. A constitutive mechanism underlies the behaviour of an entity S, that Craver describes as 'S's ψ-ing', where S is an entity (e.g. a neuron) and ψ an activity of that entity (e.g. the generation of an action potential). This phenomenon is 'constituted' by the organised entities and activities that are the components of the mechanism. According to Craver all components of the mechanism have to be parts of S.

[4] In biology, many of the levels of composition have been produced by evolution. As Simon (1962) famously argued, from an evolutionary perspective it is to be expected that complex systems possess a hierarchical organisation.

Based on this view, Craver has developed an account of multi-level mechanistic explanation in neuroscience. Levels of mechanisms are a central feature of this theory. We will now briefly present this account, and then argue for an alternative understanding of multi-level (mechanistic) explanation in biology, based on the notion of levels of composition and on our conception of biological mechanisms as causal processes or, as we prefer to call them, *causal pathways*.

Here is how Craver and Bechtel (2007, 548) explain the notion of levels of mechanisms:

> In levels of mechanisms, an item X is at a lower level than an item S if and only if X is a component in the mechanism for some activity ψ of S. X is a component in a mechanism if and only if it is one of the entities or activities organized such that S ψs.

Let us take for example a cell that divides. In this case, we have an item S (the cell) that is engaged in an activity ψ (cell division), which is explained by the mechanism for cell division. The components of this mechanism (chromosomes, centromeres and other entities and activities that together compose the mechanism responsible for cell division) are at a lower level than S's ψ-ing, i.e. the dividing cell. These component entities and activities can in turn be further decomposed into entities and activities, which are sub-components of the mechanism of cell division and thus at an even lower level, and so on. What results is a hierarchy of mechanisms for a given phenomenon; this hierarchy grounds a level-relation among all of the components, sub-components etc. of the various mechanisms in the hierarchy. These are Craver's and Bechtel's 'levels of mechanisms'.

As another example, consider Craver's (2007) analysis of the multi-level mechanism of spatial memory. According to him, this mechanism has four main levels: the level of spatial memory (the highest level), the level of spatial map formation, the cellular-electrophysiological level (which is the level of the LTP mechanisms) and lastly the molecular level. Craver claims that these four levels are best viewed in terms of levels of mechanisms, where at each level there is a decomposition of the phenomenon in terms of entities and activities the organised behaviour of which is responsible for the mechanism. So, he takes the mechanism of spatial memory to 'include NMDA receptors as components in LTP mechanisms, LTP as a component in a hippocampal spatial map mechanism, and spatial map formation as a component in a spatial memory mechanism' (2007, 266).

A first problem for the account of levels of mechanisms has to do with the conception of a 'constitutive' mechanism itself. In the mechanistic literature, the adequacy of Craver's account of constitutive relevance, which specifies what it is for something to be a component of a mechanism, has been a contested issue. The discussion has centered on whether Craver's 'mutual manipulability' theory, which makes use of Woodward's notion of ideal intervention, offers a satisfactory account of constitutive relevance (see for example Baumgartner & Casini (2017) for a critical discussion).

We have a more general objection: quite irrespective of the issue of how exactly constitutive relevance is to be understood, the 'constitutive' sense of mechanism is not a very useful notion as far as biological practice is concerned. We think that

typical and paradigmatic biological mechanisms (such as molecular mechanisms) are not to be viewed in constitutive terms, but simply as causal pathways without also having an S that ψs, i.e. a containing entity that exhibits a behaviour that is explained by the underlying mechanism. Here are two reasons to be skeptical that such containing entities exist. First, many mechanisms (e.g. cognitive mechanisms, pathological mechanisms, signal transduction pathways) are not confined within natural boundaries, and so it is not plausible to hold that mechanism components have to be parts of an object S. Second, often mechanisms can occur without the entity S the behaviour of which they are taken to 'underlie'; for example, protein synthesis can occur in cell-free systems, and so it is not plausible to take the mechanism of protein synthesis as a constitutive mechanism that explains the behaviour of an entity S (i.e. the cell). We think that these considerations show that it is better to analyse biological mechanisms in terms of causation only and not constitution, and to view them simply as causal pathways that produce the phenomena.[5]

A second problem for levels of mechanisms is that the notion of 'levels' the account incorporates is necessarily a very narrow one that we think cannot capture the use of the notion of levels in biological practice. In levels of mechanisms, X is at a lower level than Y if and only if X is a component of Y. Consider then some specific behaviour of the cell; this behaviour is 'constituted' by a mechanism, the components of which will be some (but not all) of the parts of the cell. This means that if a particular protein is not a component of a mechanism responsible for a behaviour of a particular cell, then, even if the protein is located within the cell that engages in the particular activity, it cannot be said to be at a lower level than the cell (unless every protein or other macromolecule and part of the cell is always engaging in some activity that is part of a mechanism underlying some behaviour of the cell, which however is not the case). Moreover, consider two cells (that could be located near each other) that are not both components of the same mechanism; we cannot say whether they are at the same level or not (for criticisms of levels of mechanisms, see also Shapiro, *this volume*).

While Craver is aware of these consequences of the levels of mechanisms account (see 2007, 192–3), we think that these are not just counterintuitive, but (what is more important) that they do not capture the notion of levels as it is used in biology. Both these points show that levels of mechanisms lead to a very narrow (or too local) notion of levels. But when biologists talk about the molecular, the cellular or the organismic level, they seem to use the notion in a much more global way. So, all molecules that are components of a cell are taken to be at a lower level

[5] For a detailed discussion of the arguments in this paragraph as well as of other problems of the 'constitutive' conception of mechanism see our (2022). Interestingly, in their (2021) Craver, Glennan & Povich drop the requirement that an entity S must be present when we have a mechanism. They accept that 'some mechanisms, like erosion on a riverbank or the Rayleigh scattering that makes the sky blue, are not embodied in entities; they are not mechanisms by which an entity acts, or by which a collection of entities interact' (2021, 8811). For this reason, they also drop the parthood condition of constitutive relevance.

than the cell; molecules that are not components of a particular cell, are still taken to be at a lower level than the cell; and cells that compose the muscle tissue are taken to be at the same (cellular) level with the cells that compose the neural tissue. In general, the molecular level is taken to encompass all biologically relevant molecules, irrespective of whether they are components in the same mechanism or not; similarly, the cellular level concerns the organisational level of cells and their activities, and so on for other levels. Since we think that such judgments are commonly made by biologists, what is needed is a notion of levels (such as levels of composition) that can capture these uses of the notion.[6]

While we think that levels of composition better capture the notion of levels as it is used in biology, as they can lead to a more global notion of levels than levels of mechanisms, we agree with new mechanists that many explanations in life sciences are multi-level mechanistic explanations. But since for us mechanistic constitution cannot be used to ground a hierarchy of levels of composition, we need now to explain how it is possible to have explanations spanning multiple levels if we accept the view that mechanisms are causal pathways. Let us start by presenting in some more detail our account of Causal Mechanism.

9.4 Mechanisms as Causal Pathways

The recent mechanistic literature has aimed to find a common and general notion of mechanism that is present in many different scientific fields. Such a concept is commonly thought to have both methodological value, in that looking for mechanisms is a key element of scientific method, as well as ontological significance, in that mechanisms are taken to be identifiable things in the world which underpin causal relations. So, the concept of mechanism has a double role: it is used as an explanatory concept central in scientific practice which provides understanding of how various phenomena are brought about (cf. Bechtel, 2008) as well as an ontic category which corresponds to causation-as-production, has a certain generic causal structure and can be used to construct a comprehensive metaphysics of nature (cf. Glennan, 2017).

The two roles of mechanism have not always been kept distinct. For many new mechanists, the main objective of science is to find mechanisms and to construct mechanistic explanations of the phenomena because the world consists of mechanisms, where 'mechanisms' are taken to be entities in their own right and with their own ontological blueprint. Thus, the following characterisation represents

[6] We leave it as an open question for the purposes of this paper, how exactly an account of levels in terms of mereological or compositional relations has to be strengthened in order to capture biologists' use of the concept. The account offered in Kaiser (2015) is a possible and promising option. According to Kaiser, X is on a lower level than Y iff X is a biological part of Y or 'X belongs to the same general biological kind as one or more of the biological parts of Y' (2015, 183).

a broad consensus about what a mechanism is: 'a mechanism for a phenomenon consists of entities and activities organised in such a way that they are responsible for the phenomenon' (Illari & Williamson, 2012, 120). On this characterisation, mechanisms are causal structures that involve entities, activities and a certain network of relations. Far from being neutral, this characterisation, especially by invoking activities, implies a certain generic ontic account of mechanism.

In past work (Ioannidis & Psillos, 2017, 2018; Psillos & Ioannidis, 2019a, b; see Ioannidis & Psillos (2022) for a systematic account) we have taken issue with this central tenet in the current literature about mechanisms, viz., the dual role attributed to mechanisms, and advanced and defended a novel framework for understanding mechanism as a concept-in-use in science, which we have called *Methodological Mechanism* (MM). The key tenet of MM is that mechanism should be viewed as a methodological stance, that is, as a call for a certain type of explanation. MM denies that the concept of mechanism, as a concept-in-use in science, has ontological weight. In any case, MM argues that the very issue of the ontic status of mechanisms is irrelevant to the role mechanisms play in scientific practice. MM is a deflationary position, which, without denying that searching for mechanisms is an integral and indispensable part of science, casts doubts on the claim that this search presupposes or implies that mechanism is a sui generis ontic category. Hence, MM rejects a main presupposition of much of the recent philosophical literature on mechanisms: that for the notion of mechanism, as a common and general notion present in the sciences and in particular in biology, to play its useful methodological role, it is required that it should also play its ontological role.

Far from dismissing causation, MM takes it that a mechanism for a phenomenon is a causal pathway that produces this phenomenon, as this (pathway) is described in theoretical language. We call this view *Causal Mechanism* (CM) and contend that it constitutes the correct characterisation of the general notion of mechanism found in biology. Mechanistic explanation, then, is a type of causal explanation, where the explanandum event is explained by identifying the mechanism (i.e., the causal pathway) that produces it. In particular, there are three theses that together constitute CM; these are all grounded in biological practice, which can be examined to reveal the nature and main features of mechanism as a concept-in-use. The first thesis is that mechanisms are to be identified with causal pathways as evidenced by main examples of mechanisms in biology, such as apoptosis (the mechanism of cell death) and other molecular mechanisms. The second thesis is that causal relations among the components of a pathway are to be viewed in terms of difference-making. Difference-making is important, as it is what matters in practice. On our view, and in contrast to mechanistic theories of causation, causation as difference-making is conceptually prior to the notion of a mechanism. The third thesis is that CM is metaphysically agnostic; it does not view mechanisms in ontologically loaded terms, and uses for their description the theoretical language of science.

CM, it should be noted, is thin and deflationary. It does not and cannot lead to a substantive ontological account of mechanism. It does not incorporate a distinction between entities and activities or a distinction between causal and constitutive mechanisms. Nor does it imply any 'thick' account of causation.

Causation is minimally understood as counterfactual dependence. When it comes to other 'thicker' accounts of causation CM advises agnosticism. By doing this, it goes against prevalent accounts that view mechanisms in terms of a specific (e.g. neo-Aristotelian) metaphysics. Hence, mechanisms are (simply) causal sequences, where causation is understood as a real but ontologically minimal feature of the world. According to CM, and a fortiori MM, then, mechanisms, qua causal pathways, are real things (processes) in the world, even though there is no deeper story to be told about the ontic structure and role of mechanisms. When it comes to practice, CM is enough for having a general understanding of mechanisms and their role in biology.

9.5 Multi-level Mechanistic Explanation: an Alternative View

We have claimed that biological mechanisms are causal pathways that produce the phenomena. Let us now argue for the two core claims of our account of multi-level mechanistic explanation, i.e. that levels and mechanisms are distinct notions and that levels of nature and of multi-level explanations are levels of composition.

According to Craver's and Bechtel's account, the concept of levels is dependent on the concept of mechanism, in the sense that the 'lower level than' relation is defined in terms of the componency relation, which Craver understands in terms of his theory of constitutive relevance (that involves 'mutual manipulability'—see his (2007)). What it is to be a component in a (constitutive) mechanism is thus central both for the notion of a mechanism, and for the notion of levels (of mechanisms). In contrast to this, on our view the notions of mechanism and levels are distinct. Levels of composition are grounded in mereological or compositional relations, while components of mechanisms are related by difference-making relations. For us, mechanisms are causal pathways and are not to be viewed in constitutive terms (as in Craver's and Bechtel's account); there is no relation of constitutive relevance needed in order to make sense of mechanism as concept-in-use in biology and that is also used to define hierarchies of levels. In this sense, the concept of levels is not dependent on the notion of a mechanism, and the notions are distinct. This has two main consequences. First, on our view we do not need to give an account of constitutive relevance or composition to understand what a mechanism is, since typical mechanisms in life sciences are causal pathways (more on this below). Second, since the typical sense of levels in biology is levels of composition, we do not need to understand levels in terms of mechanisms. We can thus have a more global notion of levels (based on composition), rather than the much more narrow one grounded in local mechanistic hierarchies.

The key idea of our account of multi-level mechanistic explanation is that causal pathways may contain components that belong not just to one level of composition, but to multiple ones. By saying that components of pathways belong to a certain

level of composition, we are not taking the causal relata in pathways to be objects. While the relata that stand in mereological or compositional relations are biological objects like proteins and cells, the relata in mechanisms qua causal pathways are things that can stand in causal relations. New mechanists typically think about causal relata in mechanisms in terms of entities engaging in activities, but, in giving an account of mechanism as a concept-in-use in biological practice, we prefer to remain agnostic about the metaphysics of causation and to refrain from characterising mechanisms in ontological terms. In general, we can think of the causal relata as the causally relevant properties of biological objects that are involved in the causal pathway, and which (properties) may be represented using variables. Since the causally relevant properties can be possessed by objects that belong to various levels of composition, we will say in such a case that the pathway is multi-level, or that it 'involves' objects from various compositional levels. For example, a causal pathway that involves a membrane potential as well as other molecules is a multi-level one, since the membrane potential is a property of the membrane, which is at a higher level than molecules (since it is composed of them).[7]

Multi-level mechanistic explanations, then, describe causal pathways that involve entities from multiple levels of composition. These entities (or, rather, their causally relevant properties) are nevertheless *part of the same pathway*, and in this sense are explanatorily at the same level. To be part of the same pathway, is to be a causal component of the same causal process that leads from an initial cause to an effect; and parts of the same pathway can belong, as we have seen, to various levels of composition. But all parts of a pathway, irrespective of their level of composition, are explanatorily at the same level. Since the typical notion of level in biology is, as we have argued, the compositional notion, it makes sense to call these pathways and explanations 'multi-level'.

An important feature of our account of multi-level mechanistic explanation, is that it does not require a view about how exactly to understand composition. That is, we do not need to answer the question what it is for a set of entities to compose a (biological) whole. This ontological issue does not matter for explaining how a phenomenon is produced. Moreover, multi-level mechanistic explanation, in our sense, does not commit one to (ontological) anti-reductionism. Since we view causation in terms of difference-making, for something to be a component in a pathway, it has to make a difference to the effect produced by the pathway. The capacity of a whole to be a difference-maker does not depend on whether we view it ontologically in reductionist or anti-reductionist terms; that is, it does not depend on whether it has causal powers that are in some sense over and above the causal powers of its organised components. In both cases, the whole (or the whole's

[7] Woodward (2020) has also stressed the importance of distinguishing between the relata of causal and mereological relations. But other philosophers disagree. Gillett (2016), for example, thinks that properties, powers and processes can stand in compositional relations too. New mechanists typically share this view; Craver, for example, takes the 'acting entities' that are the components of a mechanism to stand in compositional relations to the higher-level acting entity (i.e. the phenomenon for which the mechanism is responsible).

properties) can be a difference-maker and thus a component of a causal pathway. We can understand this causal role, without committing ourselves to a specific account about the ontology of composition.[8]

Let us also note that to accept the existence of levels of composition is not necessarily to reject some form of ontological reductionism or to accept some ontological version of emergence. We think that one can be a realist about compositional levels, but remain agnostic about whether (ontological) reductionism is true. Levels of composition in our sense are compatible with both a reductionist picture (as in Oppenheim and Putnam) or an anti-reductionist attitude. Similarly, we do not think that we need to view Gilbert's and Needham's talk about higher levels not being 'reducible' to lower ones in ontological terms.[9]

9.6 Some Examples of Multi-level Mechanisms

Let us consider some examples of causal pathways that contain entities from several levels of composition. A first example is the mechanism of apoptosis (see Ioannidis & Psillos, 2017, 2018). When this mechanism was first introduced by Kerr et al. (1972), it was described at the cytological level. This cytological description of the mechanism of apoptosis can be represented as follows:

condensation of nucleus & cytoplasm → budding → formation of apoptotic bodies → apoptotic bodies are phagocytosed

When Kerr et al. described apoptosis, the various molecular mechanisms that trigger the morphological changes associated with the process of apoptosis were unknown. Later, the signalling pathways that trigger the process described by Kerr et al. were uncovered. These mechanisms are described in molecular terms. For example, here is a simplified description of the so-called extrinsic pathway of apoptosis:

[8] This feature of our account is another difference with accounts such as Craver's that focus on the constitutive sense of mechanism (and 'constitutive relevance') and use it to define the level-relation. In such an account one has to specify what exactly constitutive relevance is, whereas we can have an account of multi-level explanation without saying what exactly composition is. A possible response here is that Craver's mutual manipulability account of constitutive relevance is really an epistemic criterion (see Craver et al., 2021); but then to have a complete account of constitutive relevance, something has to be said about its ontological aspect (Craver et al. argue that ontologically constitutive relevance is 'causal betweenness' which makes the overall account not very unlike Causal Mechanism). Moreover, see also Glennan (2020, 3), who claims that since 'objects are counted among the components of mechanisms … an account of corporeal composition is required to properly elucidate mechanistic constitution', and so what exactly a mechanism is.

[9] We think, however, that realism about levels of composition has enough content to rule out some ontological options. For example, vitalism (the view that some wholes have powers that are in some sense independent of their organised constituents) or dualism are ruled out; similarly, a radical eliminativist stance about higher-level entities is also ruled out, since wholes are features of reality.

Fas ligand binds to Fas receptor → adaptor protein binds to Fas receptor → procaspase-8 or 10 binds to adaptor protein → formation of DISC → activation of caspases-8 or 10 → caspases-8 or 10 activate effector caspases → destruction of proteins

As this signalling pathway is the cause of the morphological changes described at the cytological level, by combining the two descriptions, we have:

Fas ligand binds to Fas receptor → adaptor protein binds to Fas receptor → procaspase-8 or 10 binds to adaptor protein → formation of DISC → activation of caspases-8 or 10 → caspases-8 or 10 activate effector caspases → destruction of proteins → condensation of nucleus & cytoplasm → budding → formation of apoptotic bodies → apoptotic bodies are phagocytosed

The first part of the pathway (the extrinsic signalling pathway) is described in molecular terms and involves entities that belong to the molecular level; the second part of the pathway is described in cytological language and involves higher-level entities such as the cell nucleus and apoptotic bodies. This is then an example of a causal pathway containing entities from various levels of composition, i.e. a case of a multi-level mechanism.

Pathological mechanisms are also important examples of pathways that include entities from various levels of composition. Descriptions of pathological mechanisms may include reference to entities at the levels of genes and other macromolecules, cells and tissues, organs, as well as various higher-level factors such as properties of whole organisms and environmental factors such as temperatures extremes and radiation. Here is a specific example, the causal pathway of scurvy (for details see Psillos & Ioannidis, 2019b):

Citrus Fruits → Vitamin C → Scurvy

This is of course a very simplified description of the pathway. But what is important for our purposes is that this is again a pathway that contains entities from several levels of composition. In particular, (absence of) Vitamin C, the mediator, is at a lower level of organisation than the cause (which concerns the dietary habits of the organism) or the effect (which concerns the organism as a whole, e.g. feeling weak and tired, having sore arms and legs). A more detailed description of the pathway will add more levels of organisation, e.g. how the disruption of various biosynthetic pathways due to lack of vitamin C affects various tissues such as skin, gums and bones. This pathway then involves entities from several levels of composition.

The mechanism of development of type 2 diabetes is another example of a mechanism that involves many compositional levels. Diabetes is characterised by hyperglycemia (high blood sugar) due to deficiency of insulin and its symptoms include thirst and hunger, frequent urination, weight loss and feeling tired. To describe how type 2 diabetes comes about, we need to describe what happens at the level of organs, at the level of molecular mechanisms and at the level of the whole organism. Molecular mechanisms include defects in the insulin signalling pathway, which stimulates glucose uptake. Defects in this pathway may lead to

insulin resistance, in which case glucose is prevented from entering the cell. But in order to describe the pathophysiology of the disease, we need also to refer to the level of organs and tissues and to higher-level factors such as levels of glucose and insulin. The three main defects in type 2 diabetes concern the pancreas (where the pancreatic β-cells cannot produce enough insulin), muscle and adipose tissue (where due to insulin resistance glucose uptake is decreased) and the liver (where glucose production is increased); these three defects result in hyperglycemia. Importantly, type 2 diabetes has a 'natural history', beginning from an initial stage that is characterised by insulin resistance but with no further symptoms, progressing to a stage characterised by mild hyperglycemia and finally reaching a stage that requires pharmacological intervention. Insulin resistance, in particular, is influenced by both genetic and environmental (higher- level) factors, e.g. obesity and physical inactivity.[10]

9.7 Multi-level Mechanisms and Interlevel Causation

We think that these examples show that mechanisms qua causal pathways with components from various levels of composition are very common. A potential difficulty here is that such multi-level pathways involve so-called 'interlevel causation', i.e. causation between entities at different levels. While we think that scientists often make such causal claims, philosophers often view interlevel causes with suspicion. We think, however, that interlevel causation is unproblematic. In this last section we will offer some arguments in favour of this claim and contrast our account with Craver and Bechtel's analysis.

A reason why many philosophers reject interlevel causation is that this kind of causation seems to relate things that are related by mereological relations; the problem is that such relata are not spatiotemporally distinct and so they cannot stand in causal relations. Craver & Bechtel, in particular, have argued that there is no interlevel causation. For them, cases that seem to involve interlevel causes are to be viewed either as cases of constitutive relevance (if the relata are a component of the mechanism and the mechanism as a whole), or in terms of what they call 'mechanistically mediated effects'. These are 'hybrids of constitutive and causal relations in a mechanism, where the constitutive relations are interlevel, and the causal relations are exclusively intralevel' (2007, 547). An example they use is the causal claim that a virus infection caused the death of the general. While this seems to involve an interlevel (bottom-up in this case) causal claim where the effect is a property of the whole organism, what really happens, according to Craver & Bechtel, is that the virus infection leads to the disruption of various mechanisms in the organism, and ultimately produces 'the physiological conditions that constitute

[10] The case of type 2 diabetes is discussed in some more detail in Ioannidis and Psillos (2022), together with various other examples of multi-level mechanisms.

the general's death' (2007, 557). The same story can be told about cases of top-down causation: what does the causal work is a mechanism that constitutes the top-down cause and which (mechanism) produces the outcome. So, according to Craver & Bechtel, what seem like interlevel causal relations are in reality 'hybrids', where 'the putative interlevel claim is analyzed into a causal claim coupled with one or more constituency claims' (561).[11]

We reject this hybrid picture as a way to make sense of interlevel causation, as we think that interlevel causal claims are unproblematic. We have three main objections against the hybrid account. First, we think that the notion of mechanism in biology is captured by CM and thus we reject the constitutive account of mechanism which Craver and Bechtel's account presupposes. Second, we think that it is preferable methodologically to try to develop a literal construal of (what seems like) interlevel causation in biology, rather than reinterpret scientific language in part because of philosophical intuitions such as that parts and wholes cannot stand in causal relations at all. When analysing interlevel causal claims (which are ubiquitous in science) such as 'a defect in β-cells and insulin resistance cause type 2 diabetes' or 'type 2 diabetes can cause cardiovascular complications', a literal reading is to be preferred rather than reinterpreting scientists' claims in terms of the hybrid picture.

Our third reason to prefer the present account is that we do not think that interlevel causation is incoherent or conceptually problematic. We have said earlier that components of pathways are not objects, and so do not stand in mereological relations. Hence, they do not stand in mereological or compositional relations to each other. Moreover, consider examples such as the molecular pathway of apoptosis: the apoptosome is composed of various parts (e.g. Apaf-1 proteins) and so is at a higher level of composition than proteins. But it is unproblematic to say that the apoptosome has a causal role in the apoptosis pathway (that also involves entities from lower compositional levels)—similarly, it is unproblematic to say that bigger things can cause changes in smaller ones, or vice versa. Other cases are of course different; sometimes some of the objects involved in the pathway are parts of other objects. So, in the example of type 2 diabetes, the objects involved in the molecular pathways relevant to diabetes are parts of the organism that develops type 2 diabetes. However, in such cases what we have is not a synchronic causal relation from the parts to the whole; what is important to consider here is the temporal dimension. What happens in this case is that over time defects in β-cells and the insulin pathway result in changes in higher levels of organisation, and this irrespective of the fact that they are spatiotemporally contained within the organism.[12]

[11] Bechtel (2017) has argued that the account in Craver and Bechtel (2007) makes higher levels epiphenomenal, as it 'suggests a highly reductionistic picture of levels according to which causal relations that were supposed to be between entities at higher levels of organization dissolve into causal interactions at the lowest level considered' (2017, 262).

[12] Very often, causal claims that involve higher levels of organisation are to be preferred than causal claims that refer only to lower levels. As Gilbert noted, spins of quarks do not matter for what happens in the cell. A way to make this more precise is by using Woodward's notion of conditional independence (2020; see also Woodward, *this volume*). Woodward's idea is that

Thus, while our account of multi-level mechanistic explanation makes use of interlevel causal relations, we think that interlevel causation is unproblematic and in fact ubiquitous in life sciences. Moreover, our account is simpler than Craver's view that relies on the notion of levels of mechanisms and on the constitutive conception of mechanism, and simpler than Craver & Bechtel's hybrid picture of interlevel causation. In addition, it does not use the notion of mechanism to give an account of levels in biology, which as we have argued are better viewed in terms of levels of composition (that are not also 'levels of mechanisms'); the notions of mechanisms and levels are for us distinct notions. For these reasons, we think that the reconsideration of the relationship between levels and mechanisms presented in this paper can lead to a new account of multi-level mechanistic explanation in biology that is closer to scientific practice.

Acknowledgements An earlier version of this paper was presented at the workshop on the Multi-Level Structure of Reality (Israel Institute for Advanced Studies at the Hebrew University and the University of Haifa, May 2019); we thank all participants for valuable comments and suggestions, and Orly Shenker & Meir Hemmo for their invitation. We also thank audiences at the EPSA19 (Geneva, September 2019), the 6th Panhellenic Conference for the Philosophy of Science (University of Athens, December 2020), EENPS21 (University of Belgrade, June 2021) and BSPS21 (July 2021) were versions of the paper were presented, for helpful comments and criticism. SI's work has received funding from the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Innovation (GSRI), under grant agreement No 1968.

References

Baumgartner, M., & Casini, L. (2017). An abductive theory of constitution. *Philosophy of Science, 84*, 214–233.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge.

Bechtel, W. (2017). Explicating top-down causation using networks and dynamics. *Philosophy of Science, 84*, 253–274.

Craver, C. F. (2007). *Explaining the brain: Mechanisms and the mosaic unity of neuroscience*. Oxford University Press.

Craver, C. F., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*, 547–563.

Craver C, & Tabery, J. (2015). Mechanisms in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Summer 2019 Ed.). URL= https://plato.stanford.edu/archives/sum2019/entries/science-mechanisms/

'claims of downward causation (and claims of interlevel causation more generally) can be thought of as claims about the irrelevance of certain kinds of information conditional on other sorts of information—we can legitimately make claims of interlevel causation when such conditional irrelevance relations are present' (2020, 444). So, when we claim that the membrane potential V is a cause of the ionic currents and channel conductances, 'any further information about how that potential is realized in the electromagnetic forces associated with individual atoms and molecules does not matter' (2020, 444).

Craver, C. F., Glennan, S., & Povich, M. (2021). Constitutive relevance & mutual manipulability revisited. *Synthese, 199*, 8807–8828.

Eronen, M. I., & Brooks, D. S. (2018). Levels of organization in biology. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2018 Ed.). URL= https://plato.stanford.edu/archives/spr2018/entries/levels-org-biology/

Gilbert, S. F. (2010). *Developmental biology* (9th ed.). Sinauer Associates Inc.

Gillett, C. (2016). *Reduction and emergence in science and philosophy*. Cambridge University Press.

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Glennan, S. (2020). Corporeal composition. *Synthese, 198*, 11439–11462.

Illari, P., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal of Philosophy of Science, 2*, 119–135.

Ioannidis, S. & Psillos, S. (2017). In defense of methodological mechanism: The case of apoptosis. *Axiomathes*, *27*(Epistemologia 2017 Special Issue), 601–619.

Ioannidis, S., & Psillos, S. (2018). Mechanisms in practice: A methodological approach. *Journal of Evaluation in Clinical Practice, 24*, 1177–1183.

Ioannidis, S., & Psillos, S. (2022). *Mechanisms in science: Method or metaphysics?* Cambridge University Press.

Kaiser, M. I. (2015). *Reductive explanation in the biological sciences* (History, philosophy and theory of the life sciences, 16). Springer.

Kerr, J. F. R., Wyllie, A. H., & Currie, A. R. (1972). Apoptosis: A basic biological phenomenon with wide-ranging implications in tissue kinetics. *British Journal of Cancer, 26*, 239–257.

Mitchell, S. D. (2000). Dimensions of scientific law. *Philosophy of Science, 67*, 242–265.

Needham, J. (1943). *Time: The refreshing river*. Allen and Unwin.

Oppenheim, P., & Putnam, H. (1958). The unity of science as a working hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories, and the mind-body problem* (pp. 3–36). University of Minnesota Press.

Potochnik, A., & McGill, B. (2012). The limitations of hierarchical organization. *Philosophy of Science, 79*, 120–140.

Psillos, S., & Ioannidis, S. (2019a). Mechanisms, then and now: From metaphysics to practice. In B. Falkenburg & G. Schiemann (Eds.), *Mechanistic explanations in physics and beyond* (European studies in philosophy of science) (pp. 11–31). Springer.

Psillos, S., & Ioannidis, S. (2019b). Mechanistic causation: Difference-making is enough. *Teorema, XXXVIII/3*, 53–75.

Simon, H. A. (1962) [1996]. The architecture of complexity: Hierarchic systems. *Proceedings of the American Philosophical Society*, *106*, 467–482. Reprinted in his *The Sciences of the Artificial*, 3rd edition, Cambridge: MIT Press, 1996, 183–216.

Waters, C. K. (1998). Causal regularities in the biological world of contingent distributions. *Biology and Philosophy, 13*, 5–36.

Wimsatt, W. C. (1976).Reductionism, levels of organization, and the mind–body problem. ln G. Globus, I. Savodnik, & G. Maxwell, (Eds.), *Consciousness and the brain* (pp. 199–267). Plenum Press.

Woodward, J. (2020). Levels: What are they and what are they Good for? In K. S. Kendler, J. Parnas, & P. Zachar (Eds.), *Levels of analysis in psychopathology: Cross disciplinary perspectives* (pp. 424–449). Cambridge University Press.

Chapter 10
The Naturalistic Case for Free Will

Christian List

Abstract The aim of this expository paper is to give an informal overview of a plausible naturalistic case for free will. I will describe what I take to be the main naturalistically motivated challenges for free will and respond to them by presenting an indispensability argument for free will. The argument supports the reality of free will as an emergent higher-level phenomenon. I will also explain why the resulting picture of free will does not conflict with the possibility that the fundamental laws of nature are deterministic, and I will address some common objections.

10.1 Introduction

Skepticism about free will has become ever more prominent in public discourse. If one browses through the popular science literature or follows social-media coverage of the topic, one is likely to come across plenty of writings suggesting that free will is an illusion: a left-over from an outmoded, pre-scientific way of thinking that has no place in modern science. Sam Harris (2012), Jerry Coyne (2014), and Yuval Noah Harari (2016) are just three well-known writers who have made such claims. The skeptics typically appeal to a materialist worldview in which there is no place for genuine human agency and cite neuroscientific studies allegedly showing that human actions are caused not by our intentional mental states, but by physical processes in the brain and body as well as external influences. More broadly, they ask, if everything in the universe is governed by the laws of nature, and our actions are part of that universe, then how could those actions be free? And how could we legitimately be held responsible for them?

C. List (✉)
Ludwig-Maximilian University of Munich, Munich, Germany

London School of Economics, London, UK
e-mail: c.list@lmu.de

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_10

Many free-will skeptics have a noble moral motive, alongside their conviction that science is on their side: they find the present criminal justice systems in many countries unjust and wish to argue for criminal justice reform. But one can agree on the need for an overhaul of our criminal justice systems while still thinking that we shouldn't throw the notion of free will out of the window. Accepting the reality of free will is compatible with advocating criminal justice reform and supporting a more rehabilitative and less retributivist approach to punishment. And independently of its relevance to criminal justice, the idea of free will is central to our human self-understanding as agents. How, for instance, could we genuinely deliberate about which course of action to take – say, when we choose a job, a partner, or a political cause we wish to endorse – if we didn't take ourselves to be free in making this choice?

In this paper, I will outline my strategy for defending free will against the growing scepticism. Crucially, I do not proceed by denying science or watering down the definition of free will. Rather, I suggest that if we understand the lessons of a scientific worldview correctly, the idea of free will – in a fairly robust sense – is not just consistent with such a worldview but supported by it. In short, there is a naturalistic case for free will. My defence of free will is developed more fully and precisely in my recent book, *Why Free Will is Real* (2019a). The paper can be read as an informal summary of, and guide to, this defence.

I will first describe what I take to be the main challenges for free will from a scientifically informed perspective and then outline my response. And I will illustrate my strategy by zooming in on the most widely discussed challenge: the challenge from determinism. Finally, I will address some common objections.

10.2 The Challenge

Let me begin with the overall challenge. Free will can be defined, on a first approximation, as an agent's capacity to choose and control his or her own actions. Free-will skeptics argue that there is no room for this capacity in a universe in which everything is the result of physical processes. The challenge can be made more precise in terms of a general argument scheme. The skeptics typically assume that free will requires some precondition – call it property P – which might be one or perhaps all of the following:

- intentional, goal-directed agency,
- alternative possibilities among which we can choose, and
- causation of our actions by our mental states, especially by our intentions.

Then they claim that science shows that there is no such thing as property P. In particular, they argue that intentional agency, alternative possibilities, or mental causation cannot be found among the fundamental physical features of the world. Regardless of whether you consult particle physics, biochemistry, or even neuroscience, you won't get around the fact that human organisms are collections

of physical building blocks, all of which are ultimately governed by the laws of physics. And this, it seems, leaves very little room for intentional agency, alternative possibilities, and causal control over our actions. For this reason, the skeptics say, property P – whichever one of the three it is – is at best a convenient fiction of our pre-scientific way of thinking. It is not an ingredient of our physical universe. And so, since property P is required for free will, there is no free will.

Different naturalistic arguments against free will target different substitution instances for P. Some claim that intentional agency is an illusion. Intentionality does not fit into the physical universe. The idea that humans are agents with goals and purposes is a remnant from folk psychology, to be replaced by a more mechanistic understanding of the human organism as a bio-physical machine. On this picture, the traditional psychological understanding of humans as intentional agents will ultimately be replaced by a more reductionistic, neuroscientific understanding. I call this the "challenge from radical materialism".

A second set of arguments claim that if the fundamental laws of physics are deterministic, as in a mechanical clockwork, then human beings could never have any alternative possibilities to choose from. Any past state of the universe – say at the time of the Big Bang – would have been sufficient to determine everything that was going to happen thereafter. When I chose to have tea rather than coffee this morning, to give a trivial example, I could not have acted otherwise. My choice was fixed by the world's prior conditions, under the laws of nature, as was your choice to read this paper. I call this the "challenge from determinism". It is, by far, the most widely discussed challenge for free will.

A third set of arguments, finally, assert that it is illusory to think that our actions are caused by our intentions. When I act, it is my brain that makes me do it. Any consciously experienced mental state to which I might intuitively attribute my action is only an epiphenomenon accompanying the real, physical cause – a byproduct. I call this the "challenge from epiphenomenalism".

Unless we are prepared to say that intentional agency, alternative possibilities, and mental causation are not all needed for free will, the success of even just one of these arguments poses a significant challenge for free will. And on the face of it, neither intentional agency, nor alternative possibilities, nor mental causation are easy to give up as conditions for free will. Entities that don't qualify as intentional agents don't even seem to be *candidates* for the ascription of free will. Similarly, entities that never face any real choices between different alternatives don't seem to qualify as "free" either. To claim that there is free will without choices would require at least a revisionary reinterpretation of the notion of freedom. Finally, if an entity's behaviour is caused not by any intentional, mental states, but by completely sub-intentional, physical processes, then whatever this entity does can hardly be attributed to its own free will.

Furthermore, although the popular-science versions of these skeptical arguments have received much attention in public discourse, there are more academic versions too. These include, but are not restricted to, Patricia and Paul Churchland's neuroscientific arguments for "eliminativism" about intentional agency (1981, 1986), Peter van Inwagen's "consequence argument" for the incompatibility of free will and

determinism (1975), Jaegwon Kim's "causal exclusion argument" against certain non-reductive forms of mental causation (1998), and Benjamin Libet's and other scholars' experimental results on the neuronal activity underlying voluntary motor actions (1983). In sum, there appears to be a strong naturalistic case against free will.

How should we respond? One response is to conclude that there is no free will. That's what the skeptics say. I find that response unsatisfactory. My view is that we should abandon such a central tenet of our common-sense understanding of the human condition only if the arguments against it are truly compelling, and I don't think they are, as I will explain.

A second response, given by many free-will compatibilists, is to insist that free will doesn't require all of the things I have mentioned, or that it requires them only in a weaker form. For instance, one might say, it is not necessary for free will that we have alternative possibilities to choose from. What matters for free will is merely that we endorse the choices we make, not that we could have acted otherwise. Alternatively, one might redefine what it means to say that an agent "could have acted otherwise". Instead of interpreting it to mean that it was *possible* for the agent to act otherwise, one might interpret it to mean that *if* the world had been a little different than it actually was – say, the agent had tried to do something other than what he or she did – then he or she would have succeeded. We might then be able to bypass some of the challenges I have summarized. I am not convinced by such a response either, because it arguably comes at the cost of watering down the notion of free will. It's not clear that such a weakened notion can do all the work we expect the notion of free will to do, as a basis for our self-understanding as responsible agents capable of deliberating about what to do. I consider the idea that we sometimes face genuine forks in the road central to our sense of responsible agency.

My own response to the challenge is different. I concede the skeptics' starting point and accept that free will does indeed require intentional agency, alternative possibilities to choose from, and causal control over our actions. And I also concede that if we look at the world solely through the lens of fundamental physics or even that of neuroscience, we are unlikely to find agency, choice, and mental causation. But I argue that this does not show that these properties are unreal. Rather, free will and its prerequisites are emergent, higher-level phenomena. They emerge from physical processes but are not reducible to them. I will now explain this response in more detail.

10.3 Free Will as a Higher-Level Phenomenon

As noted, I accept that free will requires intentional agency, alternative possibilities among which we can choose, and causal control over our actions. I take these three properties, suitably understood, to be individually necessary for free will and jointly sufficient. That is, there is no free will without intentional agency, alternative possibilities, and mental causation, and conversely, if all three properties are present

in an entity, this is enough for free will; nothing else is needed. Characterizing free will in this way has two advantages. First, it arguably captures a relatively robust common-sense understanding of free will, in line with the "libertarian" intuitions that many people have before they encounter free-will skepticism. Second, by disaggregating free will into three properties, it clarifies what is at stake in the debate. The difficult and often emotionally charged question of whether humans have free will is replaced by a set of more tractable and somewhat less charged questions: whether humans are intentional agents, whether they have alternative possibilities, and whether their actions are caused by their mental states. This gives us a checklist of things we need to consider if we wish to find out whether there is free will. (For an earlier, similar definition of free will in terms of three properties, see Walter, 2001, p. 6.)

How, then, can we establish that humans have all three properties? The key point to note is that there are two very different ways in which we can think about human beings. We can either think of them as physical systems, consisting of gazillions of interacting particles, and insist that human behaviour is to be understood as nothing but a physical process. Or we can think of humans as not just physical but also psychological, as beings with mental states and cognitive processes that underpin their behaviour. Call the first way of thinking the "reductionistic" one, and the second the "non-reductionistic" one.

It should be clear that if we adopt the reductionistic way of thinking, we may not find support for intentional agency, alternative possibilities, and mental causation. Intentionality may not seem to be a feature of physical systems; alternative possibilities may seem to conflict with physical determinism; and mental causation seems to go against the principle that all physical events must be attributable to physical causes. So, the reductionistic way of thinking leads directly to the free-will skepticism I have described. However, the human and social sciences – from anthropology and psychology to sociology and economics – support the non-reductionistic way of thinking, which represents humans not as mere physical systems, but as agents with goals and purposes, beliefs and desires, and explains human behaviour on that basis. It would be impossible to make sense of human behaviour in its breadth and richness if we did not understand humans in this way. And this understanding, in turn, vindicates agency, choice, and mental causation as central features of human beings – features that emerge from (and "supervene" on) physical processes in the brain and body but do not lend themselves to a reductionistic description in physical terms alone.

Let me give you an analogy. Suppose someone claims that there is no such thing as unemployment. Why? Because unemployment does not feature among the properties to which our best theories of fundamental physics refer. If you consult quantum mechanics, for instance, then you won't see any unemployment. But it would be absurd to conclude from this that unemployment is unreal. It is very much a real phenomenon, albeit a social and economic as opposed to purely physical one. And of course, this verdict is supported by our best scientific theories at the relevant level, such as sociology and economics. Those theories recognize the reality of unemployment, and it features as an *explanans* and an *explanandum* in social-

scientific explanations. Like the skeptic who mistakenly searches for unemployment at the level of quantum mechanics, the free-will skeptics, I argue, make the mistake of looking for free will at the wrong level, namely the physical or neurobiological one – a level at which it cannot be found.

Free will and its prerequisites – intentional agency, alternative possibilities, and mental causation – are in the company of other emergent phenomena, from organisms and ecosystems to economies. These phenomena, too, would be hard to see if we were to look at the world solely through the lens of (say) physics or chemistry. We would see only particles and molecules, fields and forces, but no organisms, ecosystems, and economies. They are irreducibly higher-level phenomena, but that makes them no less real.

10.4 Why Are Intentional Agency, Alternative Possibilities, and Mental Causation Explanatorily Indispensable?

Let's begin with intentional agency. However much the different human and social sciences – such as anthropology on one side and economics on the other – disagree about how to explain human behaviour, the one thing they all have in common is that they take what Daniel Dennett (1987) calls an "intentional stance" towards human beings. That is, they explain human beings as agents who perceive the world and cognitively represent it, who act in pursuit of goals, and who respond to their situation in ways that are at least partly rational. Whether you consult anthropology or micro-economics, psychology or sociology, you will find this intentional mode of explanation as a common feature. By contrast, if we tried to make sense of human beings solely as heaps of interacting particles, or as complicated neural networks, we would at most be able to explain some details of the brain and body or some specific aspects of physiology and cognition – for instance, how the visual cortex implements certain perceptual tasks. We would not be able to explain the rich patterns of human behaviour in their breadth and flexibility.

To give a simple example, if I ask a taxi driver to take me to Victoria Station on one day, and I ask another taxi driver to take me to King's Cross Station on the next day, and each time I successfully reach my destination, it would be extremely hard – perhaps impossible, in practice – to explain in purely physical terms what the two events have in common. We would have to cite the incredibly complicated neural and other physical processes in each driver's brain and body as well as in the car. Contrast this with the intentional mode of explanation. Once we recognize the two taxi drivers as intentional agents who understand where I wish to go, form the intention to drive me there, and have an intelligible reason to do so, we can easily explain what's going on and make predictions on that basis. The assumption that the drivers are intentional agents is vindicated by its explanatory success. Generally, the ascription of agency to people is indispensable for a satisfactory explanation of their behaviour. This point should be fairly uncontroversial.

Next consider alternative possibilities. Just as we wouldn't be able to explain human behaviour without recognizing people as agents, so we wouldn't get very far in explaining behaviour if we didn't view people as making choices in which alternative actions are open to them. The idea that humans face choices between different options, consider them (where this can take a variety of forms ranging from quick processing to slow and careful deliberation), and select one option among the possible ones is no less important for the human and social sciences than the idea of agency itself. This means that we represent humans not as deterministic machines, but as beings for whom different courses of action are possible. I call this idea "agential indeterminism". Even in a field like decision-and-game theory in economics, which is sometimes (mis)interpreted as representing humans as nothing more than utility-maximizing automata, the notion of a decision tree with choices between several possible options is central. Having different options does not mean that they are all equally likely to be chosen. After deliberation, a decision-maker may well find some options more rational or more attractive than others.

I argue in my book that the assumption of agential indeterminism is a key presupposition of the intentional mode of explanation itself. Without that assumption, our explanations of people's behaviour in the human and social sciences would not get off the ground. My conclusion here is similar to that reached by Helen Steward (2012), who argues that the very idea of agency requires some form of indeterminism. Now one may legitimately ask whether the required agential indeterminism doesn't conflict with physical determinism. As I will explain in the next section, agential indeterminism is compatible with physical determinism – an initially surprising point which, despite sounding counterintuitive, can be established in a formally precise way.

Finally, let's turn to mental causation. Skeptics argue that it is not our conscious intentions that cause our actions, but physical states of the brain. On the theoretical side, they cite Jaegwon Kim's "causal exclusion argument" (1998), which asserts that if we attribute our actions to anything other than a physical cause, this will breach some central tenets of a scientific worldview, such as the principle that there are no physical effects without physical causes or the principle that we should not postulate more causes than strictly necessary. If a physical cause, such as a neural state of my brain, suffices to account for the movement of my arm, for instance, then we should not postulate any further mental cause. On the empirical side, skeptics cite a series of experiments conducted by Benjamin Libet and his co-authors (1983), and subsequently others, showing that, when subjects are asked to perform voluntary movements of their limbs, one can detect some preparatory brain activity – a neuronal readiness potential – before the subjects experience the conscious intention to act. Libet took this to show that our intentions are only passive byproducts of the real physical causes.

My response, which I can here summarize only briefly, is that both the theoretical and the empirical arguments against mental causation can be rebutted if we are careful in defining what we mean by "causation". If we look at how causation is understood in the special sciences, this points towards a definition of causes as systematic difference-making factors for the resulting effects. (See, for instance,

the interventionist theory of causation defended by Judea Pearl, 2000; James Woodward, 2003; and others.) Such a "difference-making" understanding of causation contrasts with a "production" understanding, which is typically assumed in epiphenomenalist arguments against mental causation (on the distinction, see Hall, 2004). Roughly speaking, on a difference-making understanding, causal regularities are counterfactual regularities that remain in place when we control for confounding factors and which can be used for effective interventions in a system. An interest-rate increase by the central bank, for instance, is a difference-making cause of a reduction in inflation. Now, the most systematic difference-making causes of human actions are often at the intentional, psychological level, not at the sub-intentional, physical one; and this remains true even if – as we may very plausibly assume – there are underlying "producing causes" (as opposed to difference-making ones) at the physical level. Versions of this claim have been defended by several scholars (see, e.g., Woodward, 2008; List & Menzies, 2009, 2017; Raatikainen, 2010; Roskies, 2012). It is our intentional, mental states that most robustly co-vary with the resulting actions, not their precise physical realizers in the brain, which are too fine-grained to qualify as difference-makers – a phenomenon that Peter Menzies and I called "realization-insensitivity".

In Libet's experiments, the neuronal readiness potentials measured prior to a subject's formation of a conscious intention are, arguably, not difference-making causes of the actions, among other things because subjects can still abort an initially intended action after the neural activity has begun. The neuronal readiness potentials are best understood as belonging to the physical implementation mechanism of voluntary action. The intentional, psychological level remains a significant site of causal regularities, all the more so when we move away from the simple motor actions studied by Libet and consider more complex actions that involve sophisticated planning. And so, the idea of mental causation remains explanatorily indispensable as well.

10.5 What Follows from This?

A skeptic might say: the present arguments only show that viewing people as intentional agents with alternative possibilities and mental causation is explanatorily useful: a convenient theoretical construct or fiction. But that doesn't imply that this is what human beings are really like. Explanatory usefulness doesn't imply reality, and the picture of human beings as choice-making agents conflicts with the more reductionistic picture given to us by the fundamental sciences. In reality, people are nothing but heaps of interacting particles.

There are two things to be said in response. First, science does not mandate adopting the reductionistic picture of human beings. To the contrary, the special sciences, from biology to the social sciences, support the alternative, non-reductionistic picture, and this picture is entirely compatible with the "physicalist" assumption that everything in the world is the result of underlying physical processes. Scientists

recognize that even if everything is grounded in physical processes, many phenomena would be impossible to explain through the lens of fundamental physics alone. Higher-level explanations, such as those we find in fields ranging from biology to the social sciences, are indispensable. The theoretical point which tends to be missed by proponents of radically reductionistic approaches is that supervenience does not imply explanatory reducibility. (For further discussion, see List, 2019b.)

Secondly, from a scientific perspective, our best guide to any questions about which entities or properties are real is given by our best scientific theories of the relevant domains. If we wish to find out whether electrons or neutrons are real, we must consult particle physics. Similarly, if we wish to find out whether the patterns of the climate are real, we must consult meteorology and climate science. This idea, defended by W. V. Quine (1977) and Arthur Fine (1984), is sometimes called the "naturalistic ontological attitude". In line with it, I suggest that if we wish to find out whether human agency, choice, and mental causation are real, we must consult our best scientific theories of human behaviour, and as noted, these theories give a positive answer.

Putting these considerations together yields an indispensability argument for free will. The argument has two premises:

Premise 1: Our best explanations of human behaviour depict humans as choice-making agents: agents with goals and purposes, alternative possibilities to choose from, and causal control over their actions. This depiction is indispensable and compatible with the rest of science.

Premise 2: If postulating certain properties or entities is indispensable in our best explanations of a given phenomenon and compatible with the rest of science, then we are (at least provisionally) warranted in taking those properties or entities to be real.

If we accept the two premises, we arrive at the following conclusion:

Conclusion: We are (at least provisionally) warranted in taking intentional agency, alternative possibilities, and causal control over one's actions to be real phenomena.

This, in a nutshell, is the core of the naturalistic case for free will. My argument is analogous to the standard naturalistic argument for realism about other properties or entities in science. Physicists are realists about particles, fields, and forces because postulating them is indispensable in our best physical theories. Biologists tend to be realists about cells, organisms, or eco-systems because postulating them is indispensable in the best theories within their domains. And psychologists, at least since the cognitive turn in their discipline, are realists about mental states and processes because postulating them is indispensable in psychological explanations. I suggest that the case for realism about intentional agency, alternative possibilities, and mental causation is no different than that for other emergent, higher-level phenomena whose reality we seldom doubt: the weather, markets, economies, and so on.

It is important to note that the naturalistic case for any ontological commitments in the sciences is always provisional. New scientific developments might render certain postulated properties or entities dispensable even when they were previously considered indispensable. In the case of free will, if new scientific developments were to undermine the first premise of my argument – that our best explanations of human behaviour depict humans as choice-making agents – then I would no longer be able to uphold my conclusion. It is good scientific practice to acknowledge this point.

Moreover, the present kind of indispensability argument for realism about some property or entity is compelling only to the extent that our best explanations of the relevant phenomena exceed a certain quality threshold. If some entity or property occurs in our best explanation of a given phenomenon, but the explanation itself is very poor, then we cannot plausibly draw any ontological conclusions from this. I will here assume, however, that the required quality threshold is met in the case of many of the psychological and social-scientific explanations that depict humans as choice-making agents.

In the next section, I will say more about why the sort of agential indeterminism that underpins the present picture of agency is compatible with physical determinism.

10.6 Indeterminism as an Emergent Phenomenon

I have argued that realism about free will is justified because the picture of humans as agents with alternative possibilities and causal control over their actions is not just compatible with science but indispensable in some of our best explanations of human behaviour. Yet, one might wonder whether the picture of humans as indeterministic, choice-making agents is compatible with a worldview in which the laws of physics could, for all we know, be deterministic. Recall that determinism means that the state of the world at any point in time fully determines the future course of events. If the world is physically deterministic, then only one sequence of events will be physically possible, given the past. Everything that will happen in the future, including all human actions, will be inevitable consequences of the past. We would therefore have to be skeptical towards any theory that implicitly or explicitly postulates indeterminism in human agency.

I want to explain why this line of reasoning is mistaken. But before I do so, I need to make a few remarks about why this is relevant. One might think that quantum mechanics, one of our best physical theories, shows that the world is indeterministic. To give a simple example, when a photon, a light particle, hits a semi-transparent mirror with a very sensitive light detector attached, there is a 50% chance that the photon will be transmitted and a 50% chance that it will be reflected. Even the entire past history of the universe appears to be insufficient to determine which of these two outcomes will occur. If this is right, then the debate about whether there could be alternative possibilities in a deterministic universe is of no practical

relevance, as our universe is indeterministic from the bottom up. However, this conclusion would be too quick. First of all, quantum mechanics is not the final word on physics. Notoriously, it has not yet been reconciled with general relativity theory, which explains phenomena such as gravity, and that theory does not share the apparent indeterminism of quantum mechanics. The jury is still out on whether a future unified theory of physics will vindicate determinism or indeterminism. Secondly, the interpretation of quantum mechanics itself is controversial, and while some interpretations, such as the standard "Copenhagen" one, take it to imply indeterminism, others do not. Rival interpretations include ones according to which some hidden variables determine which trajectory the world is on. In the case of the photon, these hidden variables would have predetermined the photon's path. This paper is not the place to discuss the interpretation of quantum mechanics. I simply want to note that the question of what physical determinism does not or does not entail is of more than hypothetical interest.

So, let me turn to the main question itself. Wouldn't physical determinism rule out the kind of agential indeterminism to which, I have suggested, our theories of human behaviour are committed?

Suppose, for the sake of argument, the world is deterministic at the fundamental physical level. How, then, could there be any indeterminism in human agency? My answer begins with the observation that the physical level is just one among many different levels at which we may describe and explain the world, and other levels, such as the chemical, biological, psychological, and social ones, are no less important from a scientific perspective. Different such levels give us different windows into reality, and it would be a mistake to consider what we see from some of those windows as less real than what see from others, especially when those windows correspond to well-confirmed scientific perspectives. When we are interested in what humans can and cannot do, the right level at which to ask this question is the level of the human and social sciences, not the fundamental physical one. This point should already be clear from what I have said so far.

But now comes a crucial point. Contrary to what is often assumed, the distinction between determinism and indeterminism cannot be drawn once and for all in a way that applies to all levels simultaneously. Rather, it is a level-specific distinction. The world may be deterministic at some levels and indeterministic at others – a point that may initially sound surprising.

To illustrate this point – as a "proof of concept" – let me introduce a toy model in which a system behaves deterministically at a micro-level and indeterministically at a macro-level (List, 2014). Consider a system which, at each point in time, is in a particular state, and where that state evolves over time in accordance with certain laws governing the system. Let's call the set of all possible momentary states in which the system could be its "state space". A "history" of the system is a possible sequence of states across time. We can think of the system's laws as constraints specifying which histories are possible and which not. For example, the possible histories could be as shown in Fig. 10.1, reproduced from List (2019a). In this example, there are six time periods, labelled $t = 1, 2, 3, 4, 5, 6$. Little dots represent states of the system, and lines from bottom to top represent histories. We

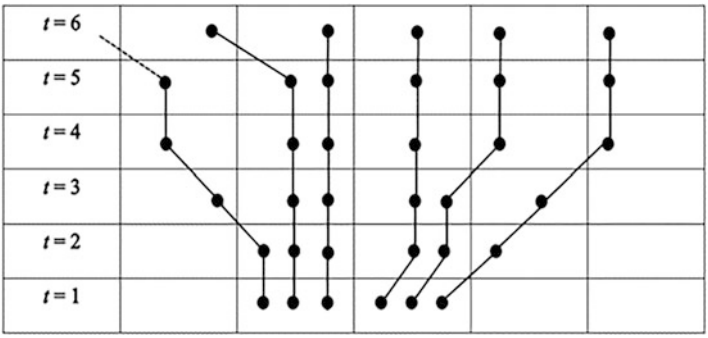

Fig. 10.1 Deterministic lower-level histories

can think of the state in the bottom row as the system's initial state, and we can think of the states along the upward-moving lines as the subsequent states. In this figure, all the possible histories are deterministic. That is to say: the initial state of each history fully determines all subsequent states; there is never any branching in any of the possible histories. We can interpret the states in Fig. 10.1 as micro-states of the system, for instance states that specify the complete configuration of all the physical particles, fields, and forces making up the system at the relevant time. Possible histories then represent the system's behaviour at a micro-level.

Now, let's suppose that we are interested in the system's behaviour at some macro-level, where the focus is not on particles, fields, and forces, but on certain macro-states. These "supervene" on the system's micro-states, but are more coarse-grained, in the sense that the same macro-state can be instantiated by different micro-states; they are "multiply realizable". An example of such a macro-state in physics is a system's temperature. Different configurations of molecules can have the same mean kinetic energy and thereby instantiate the same temperature. An example of a macro-state in psychology is a mental state such as desiring to eat chocolate and believing there is chocolate available in the kitchen. Plausibly, different neuronal configurations in the human brain could realize that same macro-state.

Formally, we can think of each macro-state as an equivalence class of micro-states, consisting of all its different possible "micro-realizers". In our example, suppose that whenever two or more different micro-states lie in the same cell of the rectangular grid in Fig. 10.1, they instantiate the same macro-state. The relevant equivalence classes are thus given by the cells. While in this toy example there are no more than three possible micro-states for each macro-state, the real systems we study in the special sciences typically admit more complex forms of multiple realizability. In principle, each macro-state could have infinitely many possible micro-realizers, and it might be infeasible to describe what they all have in common from a micro-level perspective alone. Figure 10.2, also from List (2019a), shows what our toy system looks like at the macro-level. Thick dots represent macro-states, and thick lines from bottom to top represent macro-histories.

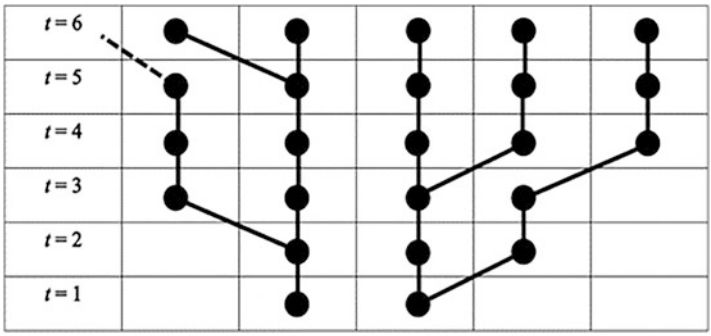

Fig. 10.2 Indeterministic higher-level histories

It is easy to see that, unlike the micro-histories, the macro-histories are not deterministic here. Regardless of the system's macro-state at time $t = 1$, several sequences of subsequent macro-states are possible: the macro-histories exhibit branching. This illustrates that macro-level indeterminism, such as the indeterminism we find in the human and social sciences, can be an emergent byproduct of micro-level determinism. More technically, the property of determinism is not preserved under changes in the level of description, such as when we move from a lower, more fine-grained level to a higher, more coarse-grained one. Crucially, all of this is entirely consistent with the higher level supervening on the lower one.

Jeremy Butterfield (2012) has expressed the same point by saying that, in a system that admits multiple levels of description, the system's micro- and macro-dynamics need not "mesh". Furthermore, one can not only go from determinism at a lower level to indeterminism at a higher one, but the reverse is also possible (for a concrete illustration, see List, 2019a, pp. 96–97). The bottom line is that indeterminism at the lower level is neither necessary, nor even sufficient, for indeterminism at the higher level. Related results were obtained by Jeffrey Yoshimi (2012) and, with a slightly different interpretational angle, by Charlotte Werndl (2009).

10.7 Some Objections

I will now address a number of common objections.

10.7.1 Isn't the Emergence of Indeterminism at a Higher Level Merely Epistemic?

The most common objection to my analysis is that, if the fundamental laws of nature are deterministic, then the appearance of indeterminism at a higher level is merely

"epistemic" – the result of our incomplete information about the system's micro-state – and so my defence of alternative possibilities fails. Even if the *macro-state* at time $t = 1$ is insufficient to determine the history of subsequent macro-states, the *micro-state* at time $t = 1$ would certainly fix all subsequent states, micro as well as macro. In consequence, what I have called "higher-level indeterminism" is an illusion due to our epistemic limitations.

However, this objection is mistaken. There are reasons for adopting an "ontic" and not merely "epistemic" interpretation of higher-level indeterminism – an interpretation which treats it as a real phenomenon. Let me sketch just a few of these reasons.

First of all, good scientific practice – in the spirit of the naturalistic ontological attitude – supports a form of pluralism about levels under which it is appropriate to take a realist attitude towards the properties at each level, provided they are explanatorily indispensable for some relevant special-science purposes and treating them as real is not incompatible with other, more fundamental commitments. Realism about higher-level indeterminism is arguably supported by this principle.

Secondly, the claim that the system's micro-state would be enough to fix all subsequent macro-states does not contradict macro-level indeterminism at all. It merely reasserts the already assumed fact that the system is deterministic at the micro-level. As an objection to macro-level indeterminism it fails, because the definition of macro-level indeterminism does not – and should not – refer to the system's micro-states. Indeed, the objection cannot even be expressed if, as is appropriate for macro-level descriptions, we refer only to macro-level facts. Macro-level indeterminism means that the system's *macro-state* at a particular time does not determine the subsequent sequence of macro-states. This definition is unambiguously satisfied in Fig. 10.2, and it is the right definition for analysing a system's macro-level dynamics, also in line with the pluralistic case for considering each level on its own terms.

Finally, we cannot assume that there is always a most fundamental level, which can somehow be treated as the privileged level for distinguishing between determinism and indeterminism "simpliciter". As Marcus Pivato and I have formally shown (2015), a scenario in which there is a bottomless hierarchy of levels, with determinism at even-numbered levels and indeterminism at odd-numbered ones, is entirely coherent, albeit hypothetical. In such a scenario, it would make no sense to speak of determinism or indeterminism "simpliciter", or to tie the distinction to any particular privileged level; after all, there is no fundamental level here. The system's indeterminism at odd-numbered levels is no more or less real than its determinism at even-numbered ones. This scenario supports the idea that the distinction between determinism and indeterminism is best understood as a level-specific distinction. And it fits nicely with the proposed ontic as opposed to epistemic interpretation of level-specific determinism or indeterminism, especially once we accept that different levels can be equally legitimate windows into reality, none of which is generally privileged over the others.

10.7.2 Isn't Agential Indeterminism Yet Another form of Randomness?

Another common objection is that, even if there is agential indeterminism, this only establishes a form of randomness or pseudo-randomness at the level of agency; but surely, this is not enough for free will.

I agree that free will requires more than randomness. First of all, however, we must not forget that agential indeterminism is only one of three requirements for free will; intentional agency and mental causation are needed too. But second and much more importantly, it would be a mistake to equate agential indeterminism with randomness. Randomness and indeterminism are not the same thing.

My analysis suggests that there are different kinds of indeterminism. Some are associated with randomness, for instance the kinds of indeterminism we find in quantum random generators or in statistical physics. In the human and social sciences, however, there is another kind of indeterminism, which is associated with option availability. In intentional explanations, we draw a crucial distinction between the options that an agent could possibly choose and those that the agent will actually choose (often for intelligible reasons). Agential indeterminism means that the set of possible options is non-singleton (meaning different courses of action are possible for the agent), not that the choice is random. And this is the kind of indeterminism required for free will, as well as the one supported by our theories of agency.

Generally, any definition of determinism is based on some underlying modal notion. Physical determinism is defined in terms of physical possibility. It is the thesis that each fully specified physical state of the world admits only one *physically possible* trajectory of future states. Physical indeterminism is the negation of this thesis. Biological determinism, if there is such a thing, would be defined in terms of biological possibility. It is the thesis that each fully specified biological state of a given system (which is more coarse-grained than any fully specified physical state) admits only one *biologically possible* trajectory of future states. Biological indeterminism is its negation. Agential determinism and indeterminism, finally, are defined in terms of agential possibility, in an analogous way. Agential possibility, in turn, is the notion of possibility used by our best theories of human agency, which I have suggested can be interpreted in terms of option availability as postulated by those theories. And while the theories represent human agents as making choices between different options – albeit perhaps not always fully rational choices – they do not represent human agents as mere randomizing devices. If we are committed naturalists, we should take this representation of human agents as choice-makers rather than randomizers at face value.

10.7.3 Are Our Best Theories of Human Agency Committed to Real Alternative Possibilities or Just to Imagined Ones?

I have adopted a realist account of agential possibility and suggested that agential possibility is the modal notion to which our best theories of human agency are committed. A critic might ask, however, whether for the sciences of human agency a purely *epistemic* notion of an agent's possibilities might suffice (a question posed, for instance, by Rosen, 2020). We might interpret the possible options postulated by our theories of human decision-making as possibilities an agent imagines, rather than as real options. On such an epistemic interpretation, as I describe it in List (2019a, p. 102), "the 'possible' options would ... be those that an agent subjectively believes to be possible, even though in reality there is only one genuinely possible option in each situation – namely, the one that ends up being chosen". This might be sufficient for explanatory purposes, without entailing any commitment to agential indeterminism.

I disagree with the suggestion that this is a good interpretation of what our theories of human action and decision-making are committed to. For a start, decision theorists – in fields like game theory and behavioural economics – do not normally think of the postulated options merely as options the agent *believes* he or she has, but as options that are *genuinely available* to the agent. Indeed, the use of concepts such as "consideration sets", "focal options", or "salient options" (which refer to *subsets* among the possible options to which a decision-maker gives particular attention) is evidence that decision theorists recognize that there is a further distinction to be drawn between options that are possible in some objective sense and options that have a certain subjective status in the decision-maker's awareness. I therefore maintain that the most literal and straightforward interpretation of the postulated options in decision theory (before we introduce notions such as consideration sets) treats them as real possibilities (from the external perspective of the decision theorist), not as imagined ones (from the internal perspective of a decision-maker).

To insist on adopting an epistemic interpretation of the modal notions of the human and social sciences while accepting an ontic interpretation of the modal notions of, say, chemistry or biology (which I think we normally do) would be to apply an unwarranted double standard. Consistently with the naturalistic ontological attitude, I propose that we should be realists about the modal notions of *all* of the special sciences, provided our commitment to them does not conflict with more fundamental commitments. And remember that my compatibility arguments show that what I say about agential possibility does not conflict with determinism at other levels, such as the physical one.

As I observe in my book, the epistemic interpretation of agential possibility "would amount to a kind of 'error theory' concerning the nature of human deliberation: agents would be systematically mistaken in thinking that they are faced with possible choices when in fact they never have more than one option" (List, 2019a, p. 103). Such an error theory might be defensible if a realist interpretation

were ruled out by other considerations, for instance if real alternative possibilities were ruled out by physical determinism. The error theory would then allow us to reconcile the *appearance* or *illusion* of choice with its lack of reality. But I have shown that alternative possibilities at the level of agency are not ruled out by physical determinism. And therefore, the error theory is not forced upon us, so we can accept the more natural realist interpretation of agential possibilities instead.

10.7.4 Isn't My Account of Free Will Vulnerable to the Problem of Present Luck?

A further objection, raised by Al Mele (2020) and Gregg Caruso (2020), is that my account of free will suffers from the "problem of present luck" or, more fully, "the problem of present indeterministic luck". Mele has argued that this problem poses a significant challenge for traditional libertarian accounts of free will, but not for standard compatibilist ones. He worries that my account may be vulnerable to it. Caruso, who is a hard incompatibilist, further worries that my account suffers from an additional problem of "constitutive luck", to which even compatibilist accounts of free will are vulnerable. I set this additional problem aside in this paper; for my response to Caruso, see List (2020).

Let me begin by explaining the problem of "present luck". Suppose we accept that free will requires a form of indeterminism in an agent's choices. Each time an agent makes a choice, say between A and B, both options are genuinely open to him or her; each is a genuine possibility.

What this means is that the agent's history up to the time in question has two possible continuations: one in which the agent does A and another in which he or she does B. Suppose, now, the agent chooses A. Does this qualify as a responsible choice? Clearly, A was not *necessitated* by the agent's prior state; B would have been equally possible. Given the same prior psychological state, it was entirely possible for the agent to do B instead. This casts doubt on whether we can genuinely attribute the choice of A to the agent rather than to a random process.

As I put the worry in List (2019a, p. 108), "If several distinct courses of action are equally consistent with the agent's full psychological state at the given time, then it is hard to see how the agent's actual action could be any more attributable to the agent than any of the other possible alternatives would have been. Why should I count as the 'author' of my action if there was nothing in my psychological state that necessitated that action?" It seems, then, that the actual choice is to be attributed more to luck than to genuinely responsible agency. This is the problem of present luck. As the libertarian Robert Kane (1999, p. 217), who recognizes the problem, puts it: "If an action is *undetermined* at a time *t*, then its happening rather than not happening at *t* would be a matter of *chance* or *luck*, and so it could not be a *free* and *responsible* action."

The critics are right that anyone who thinks that free will requires indeterminism at the time of choice must explain how an agent can be considered responsible for the resulting choices. By contrast, standard compatibilists do not face this problem. This is obvious in the case of compatibilists who think that free will does not require any alternative possibilities at all. For instance, the problem evidently does not affect those compatibilists who think that an agent is free if and only if he or she stands in an appropriate relation of authorship or endorsement to his or her actions, where this may be fully consistent with the resulting actions' being determined by the agent's character and motives. (Of course, such compatibilists may face a different kind of luck problem of their own, namely the problem of "constitutive luck", as discussed by Caruso, 2020, which is to explain how agents can be held responsible for the results of their character and motives the origins of which may have been beyond their control.) But even those compatibilists who accept an ability-to-do-otherwise requirement for free will but define it in conditional terms can avoid the "present luck" problem. Suppose I want and intend to do A, and go ahead and do A. If my ability to do otherwise simply consists in the fact that *if* I had wanted to do B instead, *then* I would have done so, the truth of this counterfactual in no way challenges the fact that I did A out of my own volition; indeed, in the actual world, in which I wanted and intended to do A, no other action would have been possible. The truth of the counterfactual does not compromise my responsibility at all.

The critics are also right that, because I take free will to require indeterminism at the level of agency, my account is more similar to traditional libertarianism from the perspective of the "present luck" problem than to standard compatibilism, and so I must confront the problem. What can I say in response?

In principle, I could respond to the "present luck" problem by drawing on and adapting one of the existing responses that have been proposed by libertarians. But my preferred response is to build on the observation that agential indeterminism is not the same as randomness. If someone has a choice between A and B, in which A and B are genuinely possible options, this does not mean that what the agent chooses is just a result of chance or randomness. As Wlodek Rabinowicz and I have argued in joint work, the genuine availability of more than one option does not preclude the intentional endorsement of one of the options by the agent, so that the chosen option stands out among the available alternatives (List & Rabinowicz, 2014).

The choice of, say, A need not be attributed to luck just because B could have been chosen instead. According to my account, what renders the agent's choice a responsible one, over which he or she can be said have control, is the conjunction of three things:

(i) According to our best explanatory theory, the agent intentionally endorses A and chooses on that basis.
(ii) Some other choice, such as B, would have been possible too, though it may not have been equally intentionally endorsed.
(iii) The agent's intention is a difference-making cause of his or her actual choice (here of A). That is, in the nearest (though perhaps not in more remote) possible worlds in which the agent has that intention, he or she makes the choice in

question, and in the nearest possible worlds in which he or she does not have that intention, he or she does not make that choice.

According to my account, all three propositions can be simultaneously true, and so it would be a mistake to conclude that the agent's actual choice is just a matter of luck.

Although a version of this response might be available to proponents of traditional libertarianism too, my account arguably has an advantage. Traditional libertarians tend not to distinguish between different kinds of indeterminism. Rather, they take one kind of indeterminism to be fundamental – typically physical indeterminism – and argue that free will requires indeterminism of that kind. Now, if that kind of indeterminism is the same as the one that supposedly underlies ordinary chance or randomness, then it is easy to see why the worries about luck may arise. From the perspective of the modal notions involved, choice, on that picture, may look similar to randomness. By contrast, a key feature of my account is the distinction between agential and physical possibility, and between indeterminism at the agential level and indeterminism at the physical one. As already noted, I argue that agential indeterminism is a *sui generis* form of indeterminism that is due to option availability, not chance. And so, on my account, it is easier than on a traditional libertarian one to differentiate agential indeterminism from chance or luck.

10.7.5 Doesn't Physicalism Entail Some Form of Reducibility of Mental Properties to Physical Ones?

My arguments for free will rest on a non-reductive physicalist view: a view according to which mental properties, while supervenient on physical properties, are non-identical to their physical realizers and causally and explanatorily significant. I reject the reductive physicalist claim that, for any mental property M, there exists a corresponding physical property P such that

(i) necessarily, M and P are co-instantiated ("equivalence in satisfaction conditions"), and
(ii) M and P can serve the same explanatory role ("substitutability for scientific explanatory purposes").

I suggest that there are combinatorial reasons as to why clause (i) is not generally satisfiable and that, even if clause (i) were satisfied, there are further conceptual reasons as to why clause (ii) is not generally satisfiable either. Critics have raised objections to both of these claims (Rosen, 2020, Kaiserman & Kodsi, 2021).

Regarding my combinatorial claim, a critic might object that if a mental property M supervenes on physical properties, then we may define a set consisting of all the possible physical configurations in which M can be instantiated; call its elements P_1, P_2, P_3, and so on. Plausibly, there are infinitely many of them. It will then be

true that M holds if and only if either P_1 holds or P_2 holds or P_3 holds, and so on. Could we then not simply identify M with the disjunction P_1 or P_2 or P_3 or ...? If we count this disjunction as a physical property, we have found a physical property that is necessarily co-instantiated with M, thereby satisfying clause (i).

My response is that the criterion for a property to count as physical is that the property is definable in the appropriate physical-level language, and I interpret this to mean that it can be defined using a finite (albeit possibly long) expression. All well-formed sentences of standard languages (natural or formal) are finite, even if there is no upper bound on the admissible length of any sentence. Now, there are strong combinatorial reasons as to why finite definability of a higher-level property in lower-level language is the exception rather than the rule.

Suppose specifically:

(a) there are infinitely many possible states in which a physical system could be (which seems reasonable to assume), and
(b) our physical-level language is countable (like all standard languages, from English to textbook logics).

Then, by assumption (a), there are uncountably many possible *sets* of states. By assumption (b), only countably many of them are describable using our given language, because the language permits only countably many expressions. This means that, in combinatorial terms, almost all sets of physical states (all but countably many) are not finitely definable in the given language. If a higher-level property such as M supervenes on physical properties, then there will certainly exist some set S consisting of all and only those possible physical states that count as instantiating M. But, as the present reasoning shows, it would be a highly exceptional case if that set S were also describable using our physical-level language. And so, finite definability of a higher-level property in lower-level terms is the exception rather than the rule.

This is an upgraded version of the classic multiple realizability argument against the reducibility of higher-level properties to physical ones (List, 2019b), which originally goes back to Jerry Fodor (1974) and Hilary Putnam (1975). If this argument is correct, then I can reject the reductive physicalist claim that any mental property is identical to some physical property.

Suppose, however, against all odds, that we can find a physical property P with which the mental property M is necessarily co-instantiated. Would I then need to concede the reductive physicalist's claim that M and P are identical? I suggest in my book that this would be too quick, because M and P may still differ in the explanatory roles they can serve. I write (on p. 69): "Because the intentional property features in semantic or logical relations, it can serve as an ingredient in intentional explanations of an agent's behaviour, e.g., by rationalizing certain actions. The neurophysiological property, by contrast, can serve at best as an ingredient in causal explanations."

A critic might worry that my analysis here either begs the question or conflates concepts and properties. As Gideon Rosen (2020) puts this point, "[t]he reductionist

who identifies M and P will say that because M and P are identical, P has whatever logical and semantic properties M has".

This is a fair objection, and I accept that I need to say more about my assumptions about property individuation. It is right that if I adopted an account of property individuation according to which any two necessarily co-instantiated properties are identical, then I would not be able to argue that M and P could differ in explanatory role even if they are necessarily co-instantiated. My response must therefore be to adopt a more fine-grained account of property individuation which allows two properties to differ in their explanatory role even in case they are necessarily co-instantiated.

While I am not here committing myself to any particular theory of property individuation (my methodology is to keep my analysis modular, so as to be able to plug in different such theories), I find it congenial to assume that properties are individuated in a suitable hyperintensional way. How exactly to do this is admittedly complicated, and I accept that there are some theoretical costs associated with individuating properties hyperintensionally. In any case, I want to emphasize that my arguments against the explanatory substitutability of M and P are subsidiary arguments. My main case against reductive physicalism rests on the combinatorial considerations summarized earlier.

10.8 Concluding Remarks

I have characterized free will in terms of three properties – intentional agency, alternative possibilities, and mental causation – and suggested that people really have those properties. Specifically, I have offered a naturalistic indispensability argument for realism about all three properties and suggested that they are ontologically on a par with many other higher-level properties we readily postulate in the special sciences. The mistake in the various forms of free-will skepticism that have recently gained popularity lies in their failure to recognize free will as a higher-level phenomenon, and in their tendency to search for free will at a lower level than the one at which it can be realistically found.

My account of free will is libertarian in one respect and compatibilist in another. It is libertarian insofar as it accepts alternative possibilities as a requirement for free will and asserts that humans really have alternative possibilities. But it is compatibilist insofar as it renders this compatible with physical determinism; hence I have proposed the label "compatibilist libertarianism".

While this may initially sound like a contradiction in terms, the consistency of my account is achieved with the help of the distinction between the agential level and the physical one. In particular, the agential indeterminism required for free will does not conflict with physical determinism. Indeed, my analysis shows that the question of whether the fundamental laws of physics are deterministic or indeterministic is completely irrelevant to the question of whether there are alternative possibilities at the level of agency. The latter question is adjudicated, not by fundamental physics,

but by our best theories of human behaviour, and these support the notion of choice between alternative possibilities as central.

There is still one important point on which incompatibilists about free will are right. Free will is not compatible with determinism *at the level of agency*. If our best theories of human behaviour were to give us a deterministic picture of human psychology, thereby refuting the sort of agential indeterminism I have defended, then this would also amount to a refutation of free will of the kind I have discussed. For the time being, however, we have good grounds for thinking that our best theories of human behaviour are not like this. They support agency, choice, and mental causation as real phenomena.

Acknowledgements This paper builds on, and summarizes ideas from, my book (List, 2019a). It draws on, and overlaps with, a series of blog posts on "The Naturalistic Case for Free Will" at the Brains Blog (August 2019), a debate with Gregg Caruso and Cory Clark (List, 2020; Caruso, 2020; Clark, 2020), and my exchange with Al Mele and Gideon Rosen at an APA symposium in January 2020. I have greatly benefitted from those scholars' criticisms, as well as from the comments of Michael Esfeld, Christian Loew, Barry Loewer, Eddy Nahmias, Kadri Vihvelin, and an anonymous reviewer. My defence of free will as a higher-level phenomenon goes back to my 2011 working paper "Free will, determinism, and the possibility to do otherwise", subsequently published as List (2014). I also build on my work with Peter Menzies on mental causation (List & Menzies, 2009, 2017), with Marcus Pivato on determinism and chance (List & Pivato, 2015), and with Wlodek Rabinowicz on alternative possibilities and intentional endorsement (List & Rabinowicz, 2014). For brevity, I omit detailed literature references here; for full references, see my book. The two closest precursors to my account of free will are Anthony Kenny's account in *Freewill and Responsibility* (1978) and Daniel Dennett's in *Freedom Evolves* (2003). Both recognize the higher-level nature of free will, but neither develops the idea in the exact same way. Another precursor is Mark Siderits's "paleo-compatibilism". He argues that "the illusion of incompatibilism only arises when we illegitimately mix two distinct vocabularies, one concerned with persons, the other concerned with the parts to which persons are reducible" (Siderits, 2008, p. 30). Finally, Sean Carroll (2016) defends a levelled understanding of free will too. Other fellow travellers in the quest to defend free will against scientific challenges include Carl Hoefer (2002), Jenann Ismael (2016), Alfred Mele (2014), Eddy Nahmias (2014), Adina Roskies (2006), and Kadri Vihvelin (2013).

References

Butterfield, J. (2012). Laws, causation and dynamics at different levels. *Interface Focus, 2*(1), 101–114.

Carroll, S. (2016). *The big picture: On the origins of life, meaning, and the universe itself*. Dutton.

Caruso, G. (2020). Why free will is not real. *The Philosopher, 108*(1), 67–71.

Churchland, P. M. (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy, 78*(2), 67–90.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified science of the mind-brain*. MIT Press.

Clark, C. (2020). The social sciences have no use for undetermined free will. *The Philosopher, 108*(1), 72–75.

Coyne, J. (2014). *What scientific idea is ready for retirement?* Edge.org

Dennett, D. (1987). *The intentional stance*. MIT Press.

Dennett, D. (2003). *Freedom evolves*. Penguin.

Fine, A. (1984). The natural ontological attitude. In J. Leplin (Ed.), *Scientific realism* (pp. 83–107). University of California Press.
Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115.
Hall, N. (2004). Two concepts of causation. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 225–276). MIT Press.
Harari, Y. N. (2016). Yuval Noah Harari on big data, Google and the end of free will. *Financial Times*, 26 August 2016.
Harris, S. (2012). *Free will*. Simon and Schuster.
Hoefer, C. (2002). Freedom from the inside out. *Royal Institute of Philosophy Supplements, 50*, 201–222.
Ismael, J. T. (2016). *How physics makes us free*. Oxford University Press.
Kaiserman, A., & Kodsi, D. (2021). Review of *Why free will is real*. *Mind, 130*(519), 987–996.
Kane, R. (1999). Responsibility, luck, and chance: Reflections on free will and indeterminism. *Journal of Philosophy, 96*(5), 217–240.
Kenny, A. (1978). *Freewill and responsibility*. Routledge.
Kim, J. (1998). *Mind in a physical world: An essay on the mind-body problem and mental causation*. MIT Press.
Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential): The unconscious initiation of a freely voluntary act. *Brain, 106*, 623–642.
List, C. (2014). Free will, determinism, and the possibility of doing otherwise. *Noûs, 48*(1), 156–178.
List, C. (2019a). *Why free will is real*. Harvard University Press.
List, C. (2019b). Levels: Descriptive, explanatory, and ontological. *Noûs, 53*(4), 852–883.
List, C. (2020). Free will: Real or illusion. *The Philosopher, 108*(1), 61–66. (article) and 76–80 (replies to critics).
List, C., & Menzies, P. (2009). Non-reductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 106*(9), 475–502.
List, C., & Menzies, P. (2017). My brain made me do it: The exclusion argument against free will, and what's wrong with it. In H. Beebee, C. Hitchcock, & H. Price (Eds.), *Making a difference: Essays on the philosophy of causation* (pp. 269–285). Oxford University Press.
List, C., & Pivato, M. (2015). Emergent chance. *Philosophical Review, 124*(1), 119–152.
List, C., & Rabinowicz, W. (2014). Two intuitions about free will: Alternative possibilities and intentional endorsement. *Philosophical Perspectives, 28*, 155–172.
Mele, A. R. (2014). *Free: Why science hasn't disproved free will*. Oxford University Press.
Mele, A. R. (2020). Free will and luck: Compatibilism versus incompatibilism. *The Monist, 103*(3), 262–277.
Nahmias, E. (2014). Is free will an illusion? Confronting challenges from the modern mind sciences. In W. Sinnott-Armstrong (Ed.), *Moral psychology. Vol. 4, freedom and responsibility* (pp. 1–57). MA (MIT Press).
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press.
Putnam, H. (1975). Philosophy and our mental life. In *Philosophical papers. Vol. 2, mind, language and reality* (pp. 291–303). Cambridge University Press.
Quine, W. V. (1977). *Ontological relativity and other essays*. Columbia University Press.
Raatikainen, P. (2010). Causation, Exclusion, and the Special Sciences. *Erkenntnis, 73*(3), 349–363.
Rosen, G. (2020). Comments on *Why Free Will is Real.* Presented at an Author-meets-Critics Symposium at the 2020 Eastern APA conference, Philadelphia.
Roskies, A. L. (2006). Neuroscientific challenges to free will and responsibility. *Trends in Cognitive Science, 10*(9), 419–423.
Roskies, A. L. (2012). Don't panic: Self-authorship without obscure metaphysics. *Philosophical Perspectives, 26*, 323–342.
Siderits, M. (2008). Paleo-compatibilism and Buddhist reductionism. *Sophia, 47*(1), 29–42.

Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
Van Inwagen, P. (1975). The incompatibility of free will and determinism. *Philosophical Studies, 27*(3), 185–199.
Vihvelin, K. (2013). *Causes, laws, and free will: Why determinism doesn't matter*. Oxford University Press.
Walter, H. (2001). *Neurophilosophy of free will: From libertarian illusions to a concept of natural autonomy*. MIT Press.
Werndl, C. (2009). Are deterministic descriptions and indeterministic descriptions observationally equivalent? *Studies in History and Philosophy of Science Part B, 40*(3), 232–242.
Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford University Press.
Woodward, J. (2008). Mental causation and neural mechanisms. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation* (pp. 218–262). Oxford University Press.
Yoshimi, J. (2012). Supervenience, dynamical systems theory, and non-reductive physicalism. *British Journal for the Philosophy of Science, 63*(2), 373–398.

Chapter 11
Physicalism: Flat and Egalitarian

Gualtiero Piccinini

Abstract Flat physicalism and egalitarian physicalism differ primarily in the following ways: (1) flat physicalism posits a fundamental level whereas egalitarian physicalism does not, (2) flat physicalism claims to be a type-identity reductionism whereas egalitarian physicalism claims to reject type-identity reductionism, (3) flat physicalism maintains that there are no levels whereas egalitarian physicalism maintains that there are levels, and (4) flat physicalism claims to be incompatible with multiple realizability whereas egalitarian physicalism claims to be compatible with multiple realizability. I argue that (1) yields an advantage for egalitarian physicalism whereas the remaining differences are due to different terminological choices and therefore are not substantive disagreements. Most importantly, egalitarian and flat physicalism agree that so-called higher-level properties are aspects of lower-level properties. Thus, modulo the appeal to a fundamental level, egalitarian and flat physicalism agree on the relation between higher-level properties and their realizers. In conclusion, physicalism should be both flat and egalitarian. It should be flat because, insofar as there are levels, they are just aspects of one and the same portion of reality. It should be egalitarian because such levels are ontologically on a par, so there is no ontological hierarchy between them.

11.1 Physicalism: Flat or Egalitarian?

Physicalism says that everything in the universe is physical. It is a metaphysical doctrine that makes sense of the success of science—including but not limited to

G. Piccinini (✉)
University of Missouri, St. Louis, MO, USA
e-mail: picciniig@umsl.edu

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_11

physics—in explaining and providing a coherent account of natural phenomena. It remains to be seen how to best formulate physicalism and whether it's true.[1]

Perhaps the greatest challenge for physicalism is to make room for what are often called *higher-level properties* or *states* of a system. (I will use "state" and "property" interchangeably, on the assumption that a state is the occurrence of a property within a system.) Special sciences describe systems in terms of all kinds of higher-level states, including *being hungry*, *being covalently bonded*, *erupting*, and so forth. Such states are not directly posited by fundamental physical theories such as classical or quantum mechanics. Although fundamental physical theories aim to tell us everything there is to say about physical systems, they need not even include the vocabulary of higher-level states. Thus, there arises the question of how so-called higher-level states relate to the states posited by fundamental physical theories.

Physicalists have traditionally debated two answers to this ontological question. *Reductive* physicalists claim that higher-level states are nothing over and above lower-level states; higher-level states reduce to lower-level states, which means that higher-level states are identical to lower-level states and that lower-level states are more fundamental than higher-level states. This is an elegant, parsimonious version of physicalism. *Nonreductive* physicalists retort that higher-level states do *not* reduce because they are *distinct* from lower-level states; higher-level states are something over and above lower-level states.

I cannot do justice to the reductionist-antireductionist dialectic here. Suffice it to say that both sides have pros and cons. The main argument for nonreductive physicalism is that many higher-level states appear to be multiply realizable, which entails that they are not identical to any lower-level state. The main argument for reductive physicalism is that, if higher-level states are distinct from lower-level states, they seem redundant—there is nothing for them to do because lower-level states are enough to cause all physical effects. Therefore, either higher-level states are epiphenomenal or they reduce to lower-level states after all.

Two independent recent developments attempt to break through this traditional dialectic: *flat physicalism* (Hemmo & Shenker, 2015, 2022; Shenker, 2017) and *egalitarian ontology* (Piccinini, 2020, 2022).[2] These two views converge on what I call the *aspect view* of levels: higher-level states are *aspects* of certain lower-level states. In other words, higher-level states are *not* additions of being over and above lower-level states—if anything, they are *subtractions* of being in the sense that, to describe higher-level states, we must omit more information about the system than we omit by describing lower-level states. The aspect view avoids the pitfalls of traditional formulations and helps turn physicalism into a clear and plausible

[1] What follows is a rough and ready characterization. For more detailed and precise surveys of the relevant background, see Bickle, 2020, Levin, 2018, Smart, 2017, Stoljar, 2017, and van Riel & Van Gulick, 2019.

[2] Although I cite Hemmo & Shenker, 2015, 2022 in Piccinini, 2020, I only learned about their work after developing my egalitarian ontology.

Table 11.1 Main disagreements between flat physicalism and egalitarian physicalism

Flat physicalism	Egalitarian physicalism
Posits a unique microstate	Does not posit a unique microstate
Reductionist	Rejects traditional reductionism
There are no levels	There are levels
Incompatible with multiple realizability	Compatible with multiple realizability

doctrine. It is an important and substantive thesis that should be included within any reasonable version of physicalism.[3]

In my previous work on egalitarian ontology, I did not emphasize physicalism as such. Instead, I formulated the view that higher-level properties are aspects of certain lower-level states as one about levels, while taking it for granted that such states are physical. I hereby remedy this previous omission by making it explicit that egalitarian ontology is a form of physicalism. In light of this, and to help compare it with flat physicalism, from now on I will refer to it as *egalitarian physicalism*.

Despite the convergence on the aspect view, there are several apparent disagreements between flat physicalism and egalitarian physicalism (Table 11.1): (1) flat physicalism posits a fundamental level—a unique microstate—whereas egalitarian physicalism does not, (2) flat physicalism claims to be a type-identity reductionism whereas egalitarian physicalism claims to reject type-identity reductionism, (3) flat physicalism maintains that there are no levels whereas egalitarian physicalism maintains that there are levels, and (4) flat physicalism claims to be incompatible with multiple realizability whereas egalitarian physicalism claims to be compatible with multiple realizability.

In this paper, I will compare and contrast the two views. I will argue that the disagreement about whether to posit a unique microstate is substantive and there are benefits in a negative answer, although this disagreement is fairly minor. The other putative differences are primarily due to different terminological choices, so there is little if any substantive disagreement on those points. Nevertheless, working through the different terminological choices will yield valuable lessons for all physicalists. I will conclude that physicalism ought to be both flat and egalitarian. It ought to be flat because, insofar as there are levels, they are just aspects of one and the same portion of reality. It ought to be egalitarian because such levels are ontologically on a par, so there is no ontological hierarchy between them.

[3] The aspect view of levels is somewhat reminiscent of the subset view of realization (e.g., Shoemaker, 2007; Wilson, 1999, 2010, 2011). For the advantages of the aspect view over the subset view, see Piccinini, 2020, 27.

11.2 Microstates

Some fundamental physical theories such as classical and quantum mechanics posit that any physical system, including the whole universe, has a microstate. A microstate is the total state of a system at any given time. Suppose that there is a unique microstate. Such a microstate is the complete state of all the microvariables that describe the actual and possible states of the system. For example, if fundamental physics could be completely formulated in terms of the positions and velocities of particles, then the exact positions and velocities of all particles in a physical system would constitute that system's microstate. If it were possible to fully describe the microstate of a system, such a description would include all the physical information that can be obtained about the system. Given the microstate of the universe, the dynamical laws posited by a physical theory could be used to infer past and future microstates of the universe to the degree they can be inferred. Following such fundamental physical theories, flat physicalism posits that any given physical system has a microstate. Implicit in flat physicalism is the assumption that the microstate is unique.

In addition to the microstate posited by some fundamental physical theories, special scientific disciplines posit macrostates. Macrostates are states of macrovariables, which describe physical systems at some level of granularity above the microstate. For example, classical thermodynamics posits macrovariables such as volume, pressure, and temperature; chemistry posits macrovariables such as oxidation, polarization, and covalent bonding; neuroscience posits macrovariables such as neuronal firing and potentiation. Macrostates are not unique—they are posited by a special scientific discipline to address a specific domain of phenomena at a specific level of granularity. Therefore, the same system may have multiple nonequivalent macrostates. For example, the very same piece of neural tissue may be described as active, as consisting of neurons some of which are firing, as consisting of molecules that are in certain chemical states, etc. In addition, macrostates may be either *complete*, including values for all the variables that determine a phenomenon at that level of granularity, or *partial*, including values for some variables that are relevant to a phenomenon while leaving other variables out.

The main thesis of flat physicalism is that a system's macrostates are *aspects* of a system's microstate. In other words, macrovariables are partial descriptions of the microstate of a system. In principle, all macrostates can be recovered from the microstate simply by omitting some information.[4] Different macrostates may capture different aspects of a system's microstate; nevertheless, they are all aspects of one and the same microstate.

[4] Needless to say, flat physicalism holds that all there is to so-called higher-level *properties* of a physical system is the system's macrostates. Therefore, since macrostates are aspects of the microstate, all there is to higher-level properties of physical systems is aspects of their microstate. I already built this corollary into flat physicalism by assuming from the beginning that a state is the occurrence of a property within a system.

The main thesis of flat physicalism is clear and compelling. It makes clear how higher-level states relate to lower-level states without positing anything more than the microstate. It also faces two challenges.

The first challenge is that even physicists working within a fundamental physical theory posit different microstates at different levels of granularity for the same system. For example, the statistical mechanics of gases is often formulated in terms of the microstate of molecules. Molecules are made of atoms, which are made of subatomic particles, some of which are made of elementary particles. More and more fine-grained microstates for a gas could be described in terms of the state of atoms, subatomic particles, or elementary particles. Although such fine-grained microstates would not be especially helpful to statistical mechanics, they would be helpful when describing atomic or subatomic phenomena.

I suppose that flat physicalists take any so-called microstate described at a level above the most fundamental physical level as something that is not a true microstate but is actually a macrostate, which is an aspect of the one and true microstate, which is the total state of a physical system at the most fundamental physical level. This response puts flat physicalism at odds with the terminology used by physicists, but that may well be a small price to pay. It also puts them at empirical risk in case it turns out that there is no fundamental physical level or scientists fail to find a unified theory that describes the fundamental level in a complete way. That is the second challenge facing flat physicalism.

Physicists are actively exploring whether there is a level of physical reality below that of the elementary particles. Such a lower level is hypothesized to explain and unify all physical forces and phenomena. If the search for such a unifying physical theory is successful, perhaps flat physicalists would say that the microstate is the state of a physical system at the level posited by the most fundamental and unified physical theory that we have or will have in the future. But there is no guarantee that there is a unique fundamental physical level. What if there isn't? What if nature has more and more structure all the way down?[5] Or what if scientists never succeed in unifying all fundamental physical forces and we are stuck with relying on multiple theories that define different nonequivalent microstates for different purposes?

Enter egalitarian physicalism. Unlike flat physicalism, egalitarian physicalism does not posit a unique microstate. Instead, egalitarian physicalism posits that the physical universe is articulated into levels and that, within any portion of reality, higher-level states are aspects of lower-level states that *realize* the higher-level states. This is very similar to the main thesis of flat physicalism. The difference is that egalitarian physicalism sidesteps the commitment to a single, unique microstate. By avoiding commitment to a unique microstate, egalitarian physicalism avoids the risk of being refuted if there is no unique microstate. It also avoids clashing with the terminology used by physicists to describe nonfundamental microstates.

[5] Metaphysicians who have raised this question include Lewis, 1991, 20; Sider, 1993; and Schaffer, 2010.

Some clarifications are needed. First, egalitarian physicalism is actually formulated in terms of both objects (particulars) and states (properties). Including objects as well, and the composition relation that occurs between objects and their proper parts, enriches our ontology in a way that helps do justice to many special sciences, because many special sciences talk about both objects and their states. In light of this, an adequate formulation of physicalism should posit both objects and properties. Nevertheless, since flat physicalism focuses primarily on states as opposed to objects, in this paper I will set objects aside and discuss only states.[6]

Second, since egalitarian physicalism does not posit a unique microstate that contains all the information about a system, it will not do to simply posit levels of more microscopic or more macroscopic states and assert that more macroscopic states are aspects of more microscopic states. The reason is that, since more microscopic states need not contain all the information about the system, there is no guarantee that more microscopic states will contain all the information needed to have more macroscopic states as their aspect. Therefore, egalitarian physicalism must include, as a separate condition, that any microstate that has a macrostate as an aspect include all the information included in that macrostate. That is why egalitarian physicalism asserts that higher-level states are aspects of their lower-level *realizers*, as opposed to aspect of a microstate. The term "realizer" adds the condition that the lower-level states include at least as much information as the higher-level states they realize.

As we have seen, flat and egalitarian physicalism agree that higher-level states are aspects of certain lower-level states. I call this the *aspect view*. The aspect view can be formulated in a theory-neutral way in terms of *portions of reality*, without invoking either a microstate or the notion of realization. This requires the assumption that we can talk about portions of reality. The notion of portion of reality is pretheoretical and informal—it's everything that occurs within a region of space over a period of time, regardless of how it is counted or described (cf. Lewis, 1991, 81, 87).[7]

The aspect view is that any description of a portion of reality describes an aspect of that portion of reality, that is, it includes *some* of the information that is contained in that portion of reality. Descriptions of a portions of reality can be more macroscopic or more microscopic depending on whether they include less or more

[6] Perhaps the distinction between objects and states breaks down in quantum mechanics; if so, since flat physicalism relies on quantum mechanics for describing the microstate, the two ontologies—with and without objects—might be reconcilable. This is a topic for future work.

[7] Physics complicates things. Instead of space and time as separate dimensions, we might consider regions of spacetime, or whatever most inclusive manifold might replace spacetime in future physics. Quantum mechanical entanglement might pose the further challenge that the state of a portion of reality might not be separable from the state of other portions of reality. The role of entanglement is probably negligible when it comes to the states studied by most special sciences, which are our primary concern here, so I disregard it here. If worse comes to worse, we can modify the notion of a portion of reality to that of everything that needs to be considered within a region of spacetime (or whatever most inclusive manifold replaces spacetime in future physics) to explain a given phenomenon.

information, respectively, about that portion of reality. Descriptions that are more macroscopic describe macrostates; descriptions that are more microscopic describe microstates. If a (relatively micro) state includes all the information that a (relatively macro) state includes plus some more information, then that macrostate is an aspect of that microstate. In the terminology of egalitarian physicalism, the macrostate is an aspect of its microstate realizer. If a microstate manages to include absolutely all the information about a portion of reality, it is the microstate *M* posited by flat physicalism. All other states are macrostates, and they are aspects of *M*.

We now have a neutral formulation of the aspect view from which both flat physicalism and egalitarian physicalism can be reconstructed. We can build on that by examining other areas of apparent disagreement.

11.3 Reductionism

Hemmo and Shenker refer to flat physicalism as a reductionist view (Hemmo & Shenker, 2019, 2022, forthcoming), whereas I reject "traditional reductionism" (Piccinini, 2020). In spite of this apparent disagreement, I will argue that flat physicalism and egalitarian physicalism agree on the relevant substantive questions.

The two main varieties of reductionism involve the epistemic reduction of one theory to another or the ontological reduction of one entity to another (van Riel & Van Gulick, 2019). What we care about here is whether so-called higher-level properties ontologically reduce to so-called lower-level properties. The clearest notion of ontological reduction is identity with a direction. In this sense, "*X* reduces to *Y*" means that $X = Y$ and *Y* is ontologically more fundamental than *X*. Thus, higher-level properties reduce to lower-level properties if and only if higher-level properties are identical to lower-level properties and lower-level properties are ontologically more fundamental than higher-level properties.[8]

In the framework we are using here, higher-level property instances are (relatively more) macro states and lower-level property instances are (relatively more) micro states. Therefore, the question of reduction transforms into the question of whether (relatively more) macro states reduce to (relatively more) micro states. Or, if there is a unique microstate, the question becomes whether macrostates reduce to the microstate. I'm going to break down this question into two:

1. are microstates ontologically more fundamental than macrostates?
2. are macrostates identical to microstates?

We've already seen that (2) has a negative answer. As flat physicalists emphasize, macrostates are not identical to microstates—rather, macrostates are (partial)

[8] Never mind that it's unclear how *Y* can be more fundamental than *X* if $X = Y$. If *X* and *Y* are the same thing, they should be equally fundamental. I discuss this tension internal to reductionism in Piccinini, 2020, 2022.

aspects of microstates. One great advantage of studying macrostates is precisely that they identify variables that make a difference to phenomena of interest without getting bogged down in the unmanageable details of microstates. Therefore, (2) is a nonstarter.

As to (1), Hemmo and Shenker point out that all there is to the relation between macro- and microstates is that the former are aspects of the latter. There is no need for further ontological relations, such as "grounding," between the two (Hemmo and Shenker, 2022). I couldn't agree more. As I've argued (Piccinini, 2020, Chap. 1), relations such as grounding and its close cousins, ontological fundamentality and priority, add nothing to our understanding of the relation between levels, and they raise unsolvable problems such as which level grounds the others (or, equivalently, which level is more fundamental than, or prior to, the others). Therefore, we should reject the ontological hierarchy generated by relations such as grounding or fundamentality between levels in favor of the egalitarian view that all levels are ontologically on a par. This is eminently compatible with flat physicalism.

The follow-up question is why Hemmo and Shenker call their view *reductionist* and what they mean by that. One reason might be that they emphasize the supposedly unique and complete microstate of physical systems and claim that so-called higher levels are just aspects of the one microstate. This may sound closer to traditional reductionism than to traditional nonreductive physicalism. Another reason is that Hemmo and Shenker adamantly reject nonreductive physicalism. This is another point on which egalitarian physicalism agrees.

Traditional nonreductive physicalism maintains that, somehow, higher-level properties are (wholly) distinct from their lower-level realizers and yet higher-level properties have their own causal powers (e.g., Fodor, 1974; Gillett, 2002, 2010; List & Menzies, 2009; Pereboom, 2002; Pereboom & Kornblith, 1991). We don't need to get too deep into the why and how of this traditional antireductionism to see that it raises more problems than it solves. An especially challenging one is the problem of causal exclusion. There is every reason to conclude that microstates causally explain all higher-level physical phenomena (cf. Papineau, 2001). If so, and if—as nonreductive physicalism maintains—higher-level properties are distinct from microstates, there is nothing left for higher-level properties to causally explain. Everything there is to do is already done by microstates (cf. Kim, 1998, 2005). That is precisely why the aspect view—that macrostates are aspects of microstates—saves both the reality and causal efficacy of macrostates.

This leads us to what Hemmo and Shenker might mean by reductionism. My guess is that, for Hemmo and Shenker, "*X* reduces to Y" means that *X* is an aspect of *Y*. If this is right, I have no objection. I just prefer to emphasize that this is not quite the traditional reductionism that traditional antireductionists oppose. As it turns out, both flat and egalitarian physicalism stand on the same side as alternatives to both traditional reductionism and traditional antireductionism.

11.4 Levels

Hemmo and Shenker (e.g., 2022) argue that everything physical is already included in microstates, and microstates are all there is. Macrostates are aspects of microstates. Therefore, there are no levels of reality. In contrast, however, I present egalitarian physicalism as an account of levels of reality—higher levels are aspects of lower levels. Therefore, there are levels after all. In spite of this apparent disagreement, again I will argue that flat physicalism and egalitarian physicalism agree on the relevant substantive issues.

I suspect that the main reason Hemmo and Shenker reject levels is that they associate levels with a hierarchical ontology in which levels are somehow (wholly) distinct from one another, there are mysterious ontological relations of grounding or fundamentality that link distinct levels, and perhaps each level is an addition of being to the levels below it. One or more of the above claims is present in most views of levels. Each of them is highly problematic. Each of them is rejected by egalitarian physicalism.

I've already rejected the views that (i) there is a grounding or fundamentality relation between levels and (ii) levels are (wholly) distinct from one another. Needless to say, I also reject the view that each level is an addition of being. Where would such addition come from, and what would it do? On the contrary, I have called higher levels *subtractions of being*. Higher levels are aspects of lower levels that we can identify and study by ignoring or subtracting away some of the details of their lower-level realizers.

This suggests that the very relation at the core of the aspect view—the *aspect of* relation—can be used to give an account of levels. Higher levels are aspects of their lower-level realizers, where realizers are just states of the system that are at least as inclusive as, and typically more inclusive than, what they realize. Although this is an account of levels, it is a flat account: all levels are just aspects of one and the same portion of reality. It is also egalitarian: there is no ontological hierarchy, no grounding or fundamentality between levels, and no addition of being to the lower level(s).

Is there any good reason to retain talk of levels, as egalitarian physicalism recommends? I can think of three. First, talk of levels is a useful shorthand for talking about different special sciences and their subject matter. Scientists themselves frequently refer to the levels of atoms, molecules, cells, organs, organisms, societies, and so forth. It's helpful to have a clear account of what that means. According to egalitarian physicalism, higher levels are as real as lower levels precisely because higher levels are just aspects of their lower-level realizers.

A second reason is that scientists often integrate descriptions and explanations at different levels. Sometimes they show that one macrovariable (Level 0) is a specific aspect of a plurality of microvariables (Level –1). Sometimes they show that the activities of a system (Level 0) are explained by the activities of its components organized in a certain way (Level –1). And so forth. A clear notion of levels

elucidates the subject matter of scientific practices that investigate the relations between levels.

A third reason is that we don't know for sure which if any level is physically fundamental. (Notice that physical fundamentality has little or nothing to do with ontological fundamentality; the physically fundamental level is just the putative level below which physics has nothing to say because, presumably, the universe has no physical articulation below it.) Since we don't know for sure which if any level is fundamental, we don't know for sure whether there is a unique microstate and at which level it is. It's hard to so much as state this question without using the notion of level. Therefore, it's better to retain the clear notion of level articulated by egalitarian physicalism so we can discuss the question of which if any level is the physically fundamental one.

In conclusion, the best thing to say about levels is that there are levels, higher levels are just aspects of their lower-level realizers, and all levels are ontologically on a par. This is a flat, egalitarian account of levels.

11.5 Multiple Realizability

The debate about multiple realizability is too complex to do it justice here.[9] For present purposes, I will discuss the following simplified question: are any higher-level states realizable in different ways by lower-level states? Hemmo and Shenker argue that multiple realizability (MR) is incompatible with physicalism—more precisely, they argue that either there is no MR or dualism is true (Hemmo & Shenker, 2015, 2022; Shenker, 2017). In contrast, I offer an account of MR that fits within egalitarian physicalism and argue that MR is common (Piccinini, 2020, Chap. 2). In spite of this apparent disagreement, again I will argue that flat physicalism and egalitarian physicalism agree on the relevant substantive issues.

Hemmo and Shanker interpret the question of MR as follows. Are there higher-level state types such that their token lower-level realizers need not have anything in common at the lower level? They have a good reason to interpret the question of MR in this way. At least some prominent proponents of MR, who pitch MR against traditional type-identity reductionism, have asserted precisely this view: that the realizers of a multiply realized state need not have anything in common at the lower level. They have even gone so far as to add that it's a big mystery how this could be so, and yet so it is (e.g., Fodor, 1997, 159).

Hemmo and Shenker's answer is unequivocally negative. Any state with a prayer of being a genuine type must have some genuine causal power. So far, this is precisely what proponents of MR maintain. But, by flat physicalism, any token state of the same type is an aspect of a microstate. Since all tokens of the same type share a causal power, the microstates of which they are aspects do have something in

[9] For a more nuanced discussion, see Piccinini, 2020, Chaps. 1 and 2.

common—the causal power that the tokens they realize share. Therefore, contra MR as interpreted by Hemmo and Shenker, the microstate realizers do have something in common, and that something in common must be present in the microstate since it's just an aspect of the microstate. I think this is absolutely correct.

But MR per se does not entail that the multiple realizers of the same higher-level state type have nothing in common. All that it requires is that there be different ways of realizing the same higher-level state type. When MR is understood in this modest way, it is compatible with the aspect view. Not only that—the aspect view provides the clearest account to date of how (modest) MR is possible. That is, MR occurs when and only when there are different ways to embed the same aspect within more inclusive microstates. In other words, MR is the very fact that occurs when the same higher-level state is contained within different types of realizer microstates. This is what my account of MR maintains.

Notice that although immodest MR would refute both traditional type-identity reductionism and flat physicalism (as well as egalitarian physicalism), modest MR only refutes traditional type-identity reductionism. Modest MR is entirely compatible with both flat physicalism and egalitarian physicalism. Therefore, Hemmo and Shenker should embrace modest MR. It's actually a feature of their flat physicalism that, implicitly, it contains a clear account of modest MR.

11.6 Conclusion: Flat, Egalitarian Physicalism

When we see through different terminological choices, it turns out that flat physicalism and egalitarian physicalism are closer than they may seem. The main issue they genuinely disagree on is whether the aspect view—the view that higher levels are aspects of certain lower levels—is best formulated in terms of the putative unique microstate of physical systems (flat physicalism) or in terms of relations between levels (egalitarian physicalism). This is a minor disagreement that should not distract us from the striking and substantive agreement between the two views.

Flat and egalitarian physicalism agree on the aspect view: any description of a portion of reality describes an aspect of that portion of reality, that is, it includes *some* of the information that is included in that portion of reality. Descriptions of a portions of reality can be more macroscopic or more microscopic depending on whether they include less or more information, respectively, about that portion of reality. Descriptions that are more macroscopic describe macrostates; descriptions that are more microscopic describe microstates. If a (relatively more) micro state includes all the information that a (relatively more) macro state includes plus some more information, then that macrostate is an aspect of that microstate.

Flat and egalitarian physicalism agree that levels are just different aspects of the same reality. Therefore, there is no ontological hierarchy between levels: higher levels are no additions of being over lower levels and there is no need to ground some levels in an ontologically fundamental level. Therefore, all levels are ontologically on a par.

Flat and egalitarian physicalism agree that the aspect view is a better account of higher-level states (and properties) than either traditional reductionism or traditional nonreductive physicalism.

Flat and egalitarian physicalism agree that immodest multiple realizability is incompatible with physicalism—there can't be higher-level physical states whose lower-level realizers have nothing in common—whereas modest multiple realizability finds its clearest account within the aspect view. That is, multiple realizability amounts to the fact that different types of microstates can share an aspect. That shared aspect is what is multiply realized.

In conclusion, physicalism ought to be both flat and egalitarian. It ought to be flat because, insofar as there are levels, they are just aspects of one and the same portion of reality. It ought to be egalitarian because such levels are ontologically on a par, so there is no ontological hierarchy between them.

Acknowledgement Thanks to Meir Hemmo, Stephen McLeod, and Orly Shenker for helpful comments on an earlier version of this paper.

References

Bickle, J. (2020). Multiple realizability. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2020 ed.). https://plato.stanford.edu/archives/sum2020/entries/multiple-realizability/

Fodor, J. A. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115.

Fodor, J. A. (1997). Special sciences: Still autonomous after all these years. *Philosophical Perspectives* 11 Mind, Causation, and the World: 149–63.

Gillett, C. (2002). The dimensions of realization: A critique of the standard view. *Analysis, 62*, 316–323.

Gillett, C. (2010). Moving beyond the subset model of realization: The problem of qualitative distinctness in the metaphysics of science. *Synthese, 177*, 165–192.

Hemmo, M., & Shenker, O. (2015). The emergence of macroscopic regularity. *Mind & Society, 14*(2), 221–244.

Hemmo, M., & Shenker, O. (2019). Two kinds of high-level probability. *The Monist, 102*, 458–477. https://doi.org/10.1093/monist/onz020

Hemmo, M., & Shenker, O. (2022). Flat physicalism. *Theoria*. https://doi.org/10.1111/theo.12396

Hemmo, M., & Shenker, O. (forthcoming). Is the Mentaculus the best system of our world? In B. Loewer, B. Weslake, & A. Winsberg (Eds.), *Time's arrow and the origin of the universe: reflections on time and chance: essays in honor of David Albert's time and chance*. Harvard University Press. forthcoming.

Kim, J. (1998). *Mind in a physical world: An essay on the mind–body problem and mental causation*. MIT Press.

Kim, J. (2005). *Physicalism, or something near enough*. Princeton University Press.

Levin, J. (2018). Functionalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2018 ed.) https://plato.stanford.edu/archives/fall2018/entries/functionalism/

Lewis, D. (1991). *Parts of classes*. Blackwell.

List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *Journal of Philosophy, 106*(9), 475–502.

Papineau, D. (2001). The rise of physicalism. In C. Gillett & B. Loewer (Eds.), *Physicalism and its discontents* (pp. 3–36). Cambridge University Press.
Pereboom, D. (2002). Robust nonreductive materialism. *Journal of Philosophy, 99*, 499–531.
Pereboom, D., & Kornblith, H. (1991). The metaphysics of irreducibility. *Philosophical Studies, 63*, 125–145.
Piccinini, G. (2020). *Neurocognitive mechanisms: Explaining biological cognition*. Oxford University Press.
Piccinini, G. (2022). An egalitarian account of composition and realization. *The Monist, 105*(2), 276–292.
Schaffer, J. (2010). Monism: The priority of the whole. *Philosophical Review, 119*(1), 31–76.
Shenker, O. (2017). Flat physicalism: Some implications. *Iyyun: The Jerusalem Philosophical Quarterly, 66*, 1–15.
Shoemaker, S. (2007). *Physical realization*. Oxford University Press.
Sider, T. (1993). Van Inwagen and the possibility of gunk. *Analysis, 53*, 285–289.
Smart, J. J. C. (2017). The mind/brain identity theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2017 ed.) https://plato.stanford.edu/archives/spr2017/entries/mind-identity/
Stoljar, D. (2017). Physicalism. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017 ed.) https://plato.stanford.edu/archives/win2017/entries/physicalism/
van Riel, R., & Van Gulick, R. (2019). Scientific reduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.) https://plato.stanford.edu/archives/spr2019/entries/scientific-reduction/
Wilson, J. (1999). How superduper does a physicalist supervenience need to be? *The Philosophical Quarterly, 49*(194), 33–52.
Wilson, J. (2010). Non-reductive physicalism and degrees of freedom. *The British Journal for the Philosophy of Science, 61*, 279–311.
Wilson, J. (2011). Non-reductive realization and the power-based subset strategy. *The Monist, 94*, 121–154.

Chapter 12
Rethinking the Unity of Science Hypothesis: Levels, Mechanisms, and Realization

Lawrence Shapiro

Abstract At least since Oppenheim and Putnam's "Unity of Science as a Working Hypothesis" (1958), many philosophers have adopted the idea that nature consists in distinct levels of organization, and that sciences, theories, or models take these levels as their subject matters. *Unity of science* requires that these sciences, theories, or models stand in a particular kind of relationship to each other. In this paper I will examine some skeptical challenges to the idea of levels and consider the conception of levels that has emerged from work on mechanistic explanation. I will then argue that instead of trying to analyze unity of science in terms of levels, it should instead be based on the realization relation. Doing so provides a coherent picture of unity of science, even if the prospects for such a unification remain dim.

12.1 General Remarks on the Unity of Science

Many philosophers of science have defended, or simply assumed, (i) a conception of nature as consisting in distinct levels of organization, and (ii) a conception of science as containing various branches – sociology, psychology, biology, chemistry, physics – that take as their subject matters these different levels of organization. Oppenheim and Putnam's "Unity of Science as a Working Hypothesis" (1958) is the *locus classicus* for "levels talk" as it is most commonly understood today. For Oppenheim and Putnam, the idea that nature's various levels of organization require examination via different sciences raised a question about whether these many sciences might, someday, be unified. According to the *unity of science hypothesis*, science is unified if and only if higher-level sciences – those that study higher levels of organization – can be *micro-reduced* to lower-level sciences.

Micro-reduction, as Oppenheim and Putnam (1958) define it, begins with the assumption that the objects constituting the subject matter of some higher-level

L. Shapiro (✉)
University of Wisconsin, Madison, WI, USA
e-mail: lshapiro@wisc.edu

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_12

science, H, are composed of parts that constitute the subject matter of some lower-level science, L. H, for instance, might contain within its universe of discourse multi-cellular organisms, and L has within its domain simply cells. H can then be micro-reduced to L if the vocabulary that H employs to describe its subject matter can be replaced with the vocabulary of L (Oppenheim and Putnam call this replacement "unity of language"). In the present example, micro-reduction entails that the multi-cellular organisms of which H speaks can be described in terms of collections of cells, for which the vocabulary of L suffices. Moreover, micro-reduction entails a "unity of laws" when the laws that describe the behavior of objects in the domain of H can be dispensed with in favor of laws that describe the behavior of the objects of which H-kinds are composed. The laws that describe how cells behave, for instance, should, if micro-reduction is possible, suffice to explain how collections of cells behave. Moreover, because the relation "is composed by" is transitive, so too is micro-reduction. Thus, if a science at level N is micro-reduced to N-1, and N-1 is micro-reduced to N-2, then the language and laws of N-2 suffice to describe and explain the subject matter of science N. Unity of science is achieved when the vocabulary and laws of the lowest-level science prove adequate to describe and explain the behavior of the objects in the domain of a highest-level science.

Just as this characterization of unity of science stems primarily from a single source, so too does its main challenge. Fodor's (1974) "Special Sciences: Or the Disunity of Science as a Working Hypothesis," is widely regarded as an incisive attack on the possibility of unity of science as Oppenheim and Putnam envisage it. In my view, this reception of Fodor's criticisms is odd, for, despite its title, Fodor's main arguments turn out not to be directed toward Oppenheim and Putnam (1958), but instead seem to target Nagel's (1961) explication of reduction (even more strangely, Fodor never cites Nagel). Fodor's main complaint is that kinds in higher-level sciences cannot be *identified* with kinds in lower-level sciences, which Nagel's approach to reduction requires. But, we have just seen, Oppenheim and Putnam's vision for unity of science requires not an identity relation between objects in the domains of different sciences, but only a relation of composition (see Shapiro, 2018). Insofar as unity of language and unity of laws follows from micro-reduction, and micro-reduction depends only on the existence of a compositional relation between objects in higher- and lower-level sciences, Fodor's criticisms miss their mark.

Of more interest to me in assessing the unity of science hypothesis are recent attacks on the very coherence of the *levels* concept – a concept whose legitimacy Fodor seems never to question and, indeed, appears happy to embrace. If no sense is to be made of levels, and if establishing the unity of science involves demonstrating a micro-reductive relationship between domains of levels of organization, then, obviously, unity of science, if possible at all, cannot be anything like what Oppenheim and Putnam imagine it to be.

Below I wish to develop a new way of thinking about the unity of science. Along the way I will, in Sect. 12.2, offer a more detailed description of how Oppenheim and Putnam view the unity of science. We shall see that their explication appears to mis-state the relata involved micro-reduction. Section 12.3 turns toward criticisms of

the notion of levels. In particular, some scientists and philosophers have questioned the *discreteness* of levels, and others have denied the presumed *correspondence* between levels of organization in nature and the sciences assigned to an investigation of these levels. In Sect. 12.4 I will present and criticize an updated characterization of levels that emerges from recent work on mechanisms. This will set the stage for Sect. 12.5, where I sketch a way of combining elements of mechanistic explanation with the relation of realization. Finally, in Sect. 12.6, I will assess the plausibility of the unity of science hypothesis in light of the apparatus developed in Sect. 12.5.

12.2 Unity of Science and Levels

Immediately following their discussion of micro-reduction, Oppenheim and Putnam introduce the idea of levels:

> As a basis for our further discussion, we wish to consider now the possibility of ordering branches in such a way as to indicate the major potential micro-reductions standing between the present situation and the state of unitary science. The most natural way to do this is by their universes of discourse. We offer, therefore, a system of reductive *levels* so chosen that a branch with the things of a given level as its universe of discourse will always be a potential micro-reducer of any branch with things of the next higher level (if there is one) as its universe of discourse (1958: 9, their emphasis).

In the ensuing clarificatory remarks, Oppenheim and Putnam emphasize that the objects in each level are to be exhaustively and exclusively composed of objects at the next lower level. The resulting picture, sometimes characterized as a "layer-cake" (e.g. Brooks, 2017), looks like this: Each layer in Fig. 12.1 marks a level of organization in nature, and each such level comprises the universe of discourse of some science, e.g. sociology, biology, cellular biology, chemistry, etc. 'Pt' is a label for the part-whole relation that holds between the objects at level N and all objects in levels higher than N. In effect, 'Pt' simply puts a different name on the compositional relation that Oppenheim and Putnam see as crucial for micro-reduction. The transitivity of Pt ensures that lower-level sciences can subsume within their universes of discourse the objects in higher levels, thus promising unified science.

I hope that the above description of levels and their role in establishing unified science, although brief, is familiar enough, or sufficiently intuitive, to provide grounds for some critical analysis. The first question to consider is whether, if Pt suffices for micro-reduction,[1] and if micro-reduction suffices for unifying the

[1] Strictly, Pt does not suffice for micro-reduction. Additionally, conditions for reduction, which are entailed by micro-reduction, must be satisfied, such as that the vocabulary in science N is not included in the vocabulary of N-1, and that the observational data that science N explains can be explained by N-1. (Oppenheim & Putnam, 1958: 5–6). These two conditions, as well as a nebulous third that concerns the systematization of theories, are criteria for reduction, which must be satisfied for micro-reduction. But we can safely assume their satisfaction in the present context.

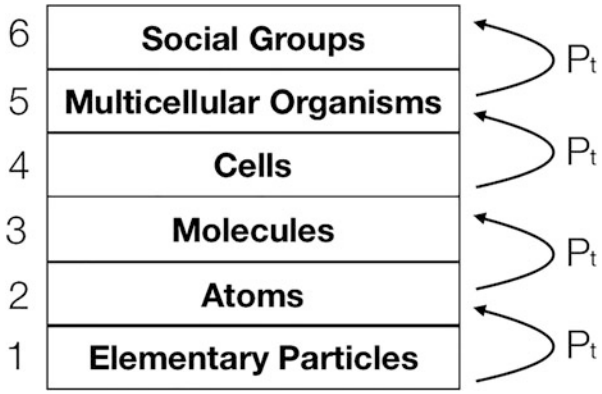

Fig. 12.1 The "layer-cake" model of levels of organization, 'Pt' is a label for the "part-whole" relation that holds between objects in distinct levels. (Adapted from Oppenheim and Putnam, 1958: 9)

sciences, the unity of science hypothesis would already have been established. What, in other words, would falsify the unity of science hypothesis given that objects in higher levels are composed from objects in lower levels? In response to this question, Oppenheim and Putnam raise the possibility that some higher-level objects may fail to be composed of lower-level objects. They consider the existence of immaterial objects, such as vital forces or entelechies, which, presumably, are not composed of objects in any of the levels that constitute nature's layer-cake. This possibility, however, they dismiss for "lack of any clear scientific meaning" (1958: 12). More seriously, the potential for emergent phenomena, recognizable in virtue of their irreducibility to "laws governing the phenomena on the level of the parts" (1958: 15), would speak against micro-reduction. However, Oppenheim and Putnam express optimism that *apparently* emergent phenomena will, as science progresses, open themselves to micro-reduction no less than ordinary phenomena. As justification for their optimism, they note the success scientists have enjoyed in synthesizing compounds, thus revealing how the properties of wholes can be determined by the interactions of their parts (1958: 15). Nevertheless, the possibility of emergent phenomena, should such optimism prove unfounded, suffices to render the unity of science hypothesis falsifiable.

A second question facing Oppenheim and Putnam's vision of unitary science focuses on the nature of the relata comprising the levels of the layer-cake. Oppenheim and Putnam speak repeatedly of the levels as consisting of objects or things. Level 5 in Fig. 12.1, for instance, consists of multicellular objects or things, i.e. organisms. The cells in Level 4 are also objects – they are the things that compose the objects in Level 5. It's important that the constituents of the levels be objects, because the relation Pt holds between objects: whole objects and the objects that are their parts.

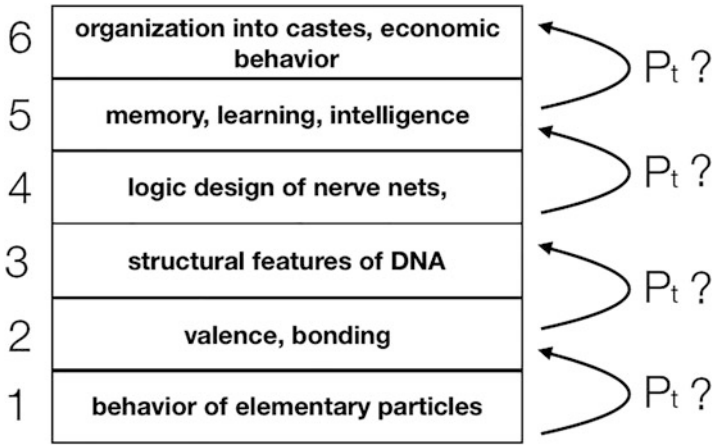

Fig. 12.2 The layer-cake model of organization in which the constituents of the layers are phenomena, processes, and properties

Yet, despite this commitment to object-talk, when Oppenheim and Putnam turn to a discussion of evidence for micro-reduction, objects and their parts fade from view. We get instead an analysis of how *phenomena*, like caste-based social structures among insects, result from the effect of social hormones on individuals. We are told of *processes*, like cellular duplication and mutation, that can be explained in terms of other processes involving DNA (O&P: 20–23). Similarly, we're told how *properties*, like "the high fluidity of water, the elasticity of rubber, and the hardness of diamond" (O&P: 22) can be micro-reduced to the atomic level. Indeed, if the layer-cake in Fig. 12.1 were re-designed to reflect the discussions of micro-reduction that Oppenheim and Putnam actually present, its appearance would be something like in Fig. 12.2:

This new layer-cake tempts one to stories like the following: the economic behavior of people (how they make investment decisions, say) can be explained in terms of their learning histories (and other psychological traits), which can be explained in terms of activities in neural networks, which can be explained in terms of the behavior of neurons, which depend on how DNA has designed the neurons, which is explained in terms of chemical reactions, and so on.

But if this is how micro-reduction is to go – and it resembles the discussions of "successful" micro-reductions that Oppenheim and Putnam offer – then micro-reduction does not rest on the relation Pt after all. Consider just the first step in the story above. The economic decisions a person makes can be explained by the person's learning history (as well as other psychological processes). But economic decisions are not composed of learning histories. The learning histories do not stand to economic decisions as a part stands to a whole – as baking soda, say, stands to a pancake. Similarly, the account for how connection weights in a neural network explains learning does not depend on learning being a composite object of which the

connection weights are a part. A neural network with some given set of properties is not a *part* of an agent's psychology.

In short, as the question marks next to the 'Pt' labels in Fig. 12.2 indicate, the constituents in the layers of this new cake seem not to stand in compositional relations to each other. The relation might better be described as a kind of supervenience. Economic decisions *supervene* on psychological processes, which *supervene* on brain processes, which *supervene* on neuronal behavior, and so on. This is plausible: two individuals with identical neuronal behavior will be identical with respect to their brain processes; and if identical with respect to their brain processes, they will be identical with respect to their psychological processes, and so on. Rendered this way, we can make better sense of how the events in lower levels can explain events in higher levels. The explanation will appeal to relations of supervenience, not composition.

Accepting that the relata of micro-reductions stand in supervenience, rather than compositional, relations to each other remains consistent with Oppenheim and Putnam's concerns about immaterial and emergent phenomena. If a theory like property dualism turns out to be true, and psychological properties do not supervene on physical properties, this would prevent the unity of science. Emergent properties must be treated slightly differently, for, on most accounts (Kim, 1998a), emergent properties do supervene on lower-level physical properties: any two individuals who are identical with respect to their lower-level properties will be identical as well with respect to emergent properties. However, unlike ordinary cases of supervenience, where the properties in the supervenience base can provide an explanation for the powers of the supervening properties, this is not so with emergent properties. Full understanding of the behavior of a supervening base still leaves obscure the behavior of the emergent property. For this reason, the existence of emergent properties would preclude unity of laws, and so would prevent micro-reduction.

12.3 Troubles with Levels

But perhaps the world is not organized as depicted in either Figs. 12.1 or 12.2. Or perhaps it is, but sciences do not correspond to distinct levels. In either case, unity of science cannot proceed as Oppenheim and Putnam would have hoped. If levels of organization do not exist, then sciences do not take as their universes of discourse the contents of levels, and so unification, which is supposed to involve the subsumption of higher-level sciences – those that study higher levels of organization – by lower-level sciences, could not occur. Similarly, if levels of organization did exist, but sciences did not dedicate themselves to the investigation of discrete levels, then the very idea that some sciences are at lower- or higher-levels than others no longer makes sense. Potochnik and McGill (2012), following Guttman (1976), deny that nature is organized into distinct levels of organization. Craver (2007) rejects the idea that for each level of organization there corresponds a distinct science.

With respect to the viability of levels of organization, Guttman reviewed a variety of examples in which nature appears not to observe layering of the sort that appears in either Figs. 12.1 or 12.2. On these models, each layer purports to contain an exhaustive and exclusive collection of parts, the interactions between which suffice to explain the behavior of the next higher level of organization. However, Guttman notes that ecosystems are composed of not just populations, but also molecules of waste material, "which may be food materials for important organisms of the system" (1976: 112). This observation clashes with Oppenheim and Putnam's vision of nature insofar as something at a "higher-level", an ecosystem, consists in interacting parts, such as individuals and molecules, that span a variety of lower-levels. Similarly, Guttman points out that "tissue is built of cells, but it is also built to a large extent of macromolecules that bind cells together. If blood is considered a tissue, then it consists of cells, macromolecules, and small molecules all together" (1976: 112). This again spells trouble for Oppenheim and Putnam's levels, for it puts the lie to the idea that a higher-level kind – tissue – can be decomposed into kinds exclusively at the next level down, which in turn can be decomposed into kinds at the next lower level down, and so on. Composing organic tissue are not just cells, but things like macromolecules, where these macromolecules needn't themselves be components of cells. As Potochnik and McGill summarize Guttman's critique of levels: "it is certainly not the case that every whole is composed of only parts at the next lower level" (2011: 127).

On reflection, Guttman's point is very obvious. Consider something with a high-level of organization, like a university. What are the parts of a university at the "next level down," whose interactions explain the organization of the university? The university's organization is a product of interactions between other organizations, e.g. a faculty senate, and single individuals, e.g. a provost, and objects, like money, whose role is determined by social factors, and even larger organizations, like state and federal governments, that shape the university through legislation or funding decisions. No sense can be made of the idea that these determiners of a university's organization belong to a single level, or, in many cases, that they stand in compositional or supervenience relations to each other. Surely Guttman is correct that the organization of ecosystems, organisms, and tissue no more readily conforms to the layer-cake model than does a university.

As Guttman doubts the possibility of imposing upon nature a layer-cake structure, Craver distrusts Oppenheim and Putnam's belief that sciences arrange themselves to correspond to distinct hierarchical levels of organization. Craver's point might be taken as a corollary to Guttman's in the following sense. If it is true that an explanation of the organization of something like an ecosystem requires that one examine interactions between groups (e.g. populations), individuals (e.g. predators), microscopic organisms (e.g. viruses), climate, chemicals (e.g. pheromones), and so on, then the study of ecosystems must draw on numerous theories and models, including those deriving from evolutionary biology, zoology, sociology, microbiology, chemistry, and so on. In such a case, the idea that there exists a single science, e.g. ecology, that can explain the behavior of ecosystems in a single

proprietary vocabulary, seems hopeless.[2] The same point is readily apparent when considering a phenomenon like clinical depression. Such a condition might result from any number of factors: childhood trauma, imbalances in serotonin levels, brain injury, alcohol abuse, and so on. There may be no single science equipped to offer a full explanation of clinical depression, or there may be many sciences capable of offering distinct insights into its manifestation. More generally, as Craver says, "[s]ingle fields increasingly reach across multiple levels of nature, and different fields often approach items at the same level of nature from different perspectives" (2007: 176). Together, these points reveal the implausibility in Oppenheim and Putnam's belief that sciences contain within their domains of discourse objects at a single level of organization, or that objects at a single level of organization open themselves to inspection from only one science.

How serious are Guttman's and Craver's worries for the prospect of unifying science? At the very least, they reveal the inaccuracy of the layer cake model of nature, and the error in supposing that subjects matters exclusive to each layer must be the targets of single sciences. Nature is not, after all, organized into rigidly partitioned levels of organization, and explanations of natural phenomena typically rely on theories, models, and laws from a range of scientific fields. We can go further and say that the composition relation, Pt, on which micro-reduction depends, turns out not to capture the only important feature of nature. Doubtless, composite things are composed by parts – that's tautological – but Guttman's point, in abstract terms, is this. The parts that determine the behavior of a composite may do so in virtue of interacting with other composites, or with parts of parts, thus ruining the neat ordering that the layer cake model seeks to impose on nature. And, even if, as I suggested, we replace the layer cake in Fig. 12.1 with that of Fig. 12.2, and replace Pt with a different sort of relation, e.g. supervenience, the same problem arises. A composite, C, may supervene on a collection of parts, P, but there is no reason to expect that P consists in kinds of just one level, so again we are left with less a layer cake and something more like a fruitcake, with objects of different sizes and complexity – walnuts, dates, raisins, candied fruits – spread throughout. All this suggests that science cannot be unified in anything like the sense that Oppenheim and Putnam imagined, and that talk of levels, if to be preserved, must be grounded in some way other than by appeals to relations like Pt or supervenience.

This last remark sets the stage for Craver's mechanistic account of levels. Craver's understanding of levels departs significantly from Oppenheim and Putnam's, and is not intended as an approach to unifying the sciences. However, we'll see, it suggests a picture for unifying sciences that I'll take some steps toward developing in Sects. 12.5 and 12.6.

[2] Craver illustrates his point with a discussion of spatial memory. An explanation of how spatial memory works will draw on "anatomy, biochemistry, computational neuroscience, electrophysiology, molecular biology, neuroanatomy, pharmacology, psychiatry, and experimental psychology" (2007: 176).

12.4 Levels in Mechanisms

In recent years, a number of philosophers have offered detailed accounts of mechanisms, seeking to answer questions concerning their ontology, how they are to be individuated, how to make sense of the relations they bear to their parts, what it means to explain their activities, how causal claims are to be understood in a mechanistic context, and so on (Bechtel, 2008; Glennan, 2017; Machamer et al., 2000). Fortunately, despite differences in how these questions are to be answered, a rough consensus has emerged regarding what mechanisms are. A foundational paper on the topic, Machamer et al. (2000), defines 'mechanism' as follows: "Mechanisms are entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions" (2000: 3). As this definition suggests, in characterizing a mechanism, one must first identify a phenomenon that one wishes to understand – a product – and then proceed to identify the organization of entities and activities that results in the production of the phenomenon. As an example, Machamer, Darden, and Craver mention DNA replication. This phenomenon is the result of a mechanism: "the DNA double helix unwinds, exposing slightly charged bases to which complementary bases bond, producing, after several more stages, two duplicate helices" (2000: 3). This simple explanation of DNA replication mentions entities, such as helices and charged bases, and activities, such as unwinding and bonding. The explanation succeeds in virtue of describing how these entities, through their activities and organization, produce the phenomenon of interest: DNA replication.

In the present context, the interest in mechanisms stems from the fact that the entities and activities of which mechanisms are composed are themselves often mechanisms, composed of entities and activities that, due to their organization, result in a product that then contributes to the product of the larger mechanism in which they are contained. The bases, for instance, which are components in the mechanism for DNA replication, are productive in virtue of how the acting entities that compose them – nucleotides, connected to each other via hydrogen bonds – are organized. And the nucleotides too can be understood as mechanisms consisting of organizations of nucleosides and phosphate groups.

The nesting of mechanisms within mechanisms instantly suggests a hierarchy that might provide a new way of thinking about levels. And so, for Craver (2007, 2015), and Craver and Bechtel (2007), it does. From the outset, however, Craver is explicit that his mechanism-inspired levels are nothing like Oppenheim and Putnam's levels (or Wimsatt, 1976). A mechanism-based analysis of levels "eschews the idea that levels are monolithic strata in the structure of the universe, with proprietary causal laws and forces" (2015: 23), as Oppenheim and Putnam (and Wimsatt, 1976) would have insisted. What, then, are levels within mechanisms, and what value does the introduction of "level-talk" play within scientific research?

Figure 12.3 below clarifies the conception of levels that mechanisms offer.

Intuitively, the components of S, the various X_is, are at a lower level than S; and similarly, the components any given X_i are at lower level still. But so far,

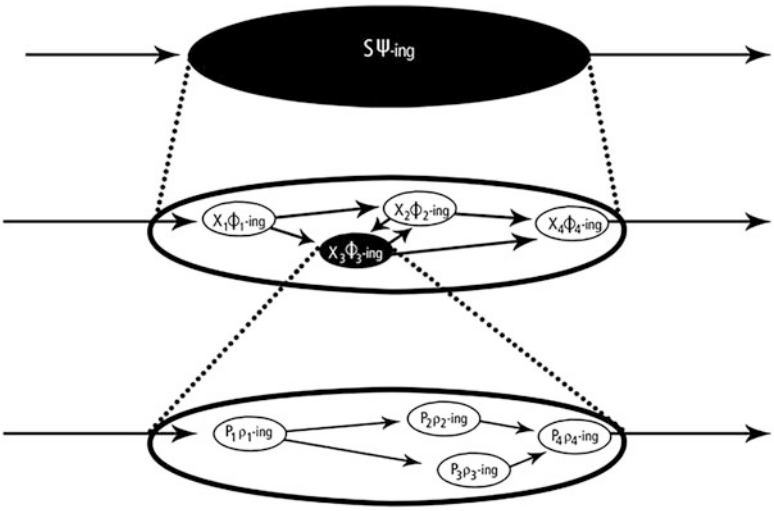

Fig. 12.3 A mechanism, S, which Ψs, is composed of entities, X_{is}, that produce Φ-ings, which are themselves mechanisms composed of P_{is} that produce ρ-ings. (From Craver 2007: 189)

there appears to be nothing more to the idea of levels than what talk of parts and wholes affords. Why should we accept that each of the X_is is at the same level simply because each is a part of S? Indeed, one might see Guttman's objections to Oppenheim and Putnam as showing that parts do not belong to the same level merely in virtue of being parts. The cells and the molecules that bind them are both parts of biological tissue, but, intuitively, the cells and the molecules are at different levels. Making sense of levels demands criteria by which we might judge when two entities are at the same level and when they belong to different levels.

Craver, however, vacillates about what these criteria are, or whether they're even necessary. He says first that "[s]ome component, X's Φ-ing, is at a lower mechanistic level than S's Ψ-ing if and only if X's Φ-ing is a component in S's Ψ-ing, that is, if and only if X's Φ-ing is a relevant spatiotemporal part of S's Ψ-ing" (2015: 17).[3] This definition asserts only that a component of a mechanism is at a lower level just in case it is a component. In other words, this definition draws an equivalence between the part-whole hierarchy and the levels hierarchy. How does such an analysis of levels differ from Oppenheim and Putnam's which, as we saw, also defines levels in terms of a part-whole relation? As I see it, Craver and Oppenheim and Putnam have approached matters from different directions. Oppenheim and Putnam take for granted that nature consists in distinct levels of

[3] I find this locution cumbersome and will continue to use single letters to refer to mechanisms and components. Craver's motivation for labeling mechanisms and components as he does, as S's Ψ-ing and X's Φ-ing, is to ensure that one does not lose sight of the fact that mechanisms are defined in terms of what they do (2015: 17).

organization, and that claims like "X and Y are at the same level of organization," or "X and Y are at different levels of organization" are sensible. Whether science can be unified then depends on whether the various levels of organization can be tied together in some way, and, the thought goes, if the levels are related to each other in terms of composition, such tying together may be possible.

Craver, on the other hand, simply asserts that components differ in level from the wholes of which they are parts, but this leaves obscure what role levels are playing for Craver. For Putnam and Oppenheim, levels are a fact and the unity of science a hypothesis regarding their relationship. But, unless Craver provides a means to identify kinds as existing at the same or different levels, his mechanistic levels are not levels at all. The levels concept entails that some things are at the same level and other things not.

This point becomes clearer when examining Craver's efforts to say something substantive about levels. He says: "X's Φ-ing and S's Ψ-ing are at the same level of mechanisms only if X's Φ-ing and S's Ψ-ing are components in the same mechanism, X's Φ-ing is not a component in S's Ψ-ing, and S's Ψ-ing is not a component in X's Φ-ing" (2015: 19). But notice that this statement provides only a necessary condition for sameness in level, not a sufficient condition (Eronen, 2015 also makes this point, as well as many of those that follow). That is, it tells us when two components fail to be at the same level: they cannot be in different mechanisms and they cannot be components of each other. But suppose we want to know the conditions for determining when two components are at the same level. By analogy, if we want to know whether two girls are both Girl Scouts, it's hardly helpful to be told that they both have heads, assuming that having a head is a necessary condition for being a Girl Scout. We want to know what *makes* someone a Girl Scout, not what prevents someone from being a Girl Scout.

To see the importance of this question, consider some of the possibilities that the necessary condition for sameness-in-level leaves open. Applied to Fig. 12.3, the condition tells us that two token Xs are at a different level than S, because one of the necessary conditions for sameness in level has been violated: the Xs are components of S. But now look at Fig. 12.4:

The two identical mechanisms, both Ss, might be tokens of the same type of spark plug in an engine, or neuron in a brain. Because they are identical, their components will also be identical. But, the circled components belong to distinct mechanisms. Thus, Craver's necessary condition for sameness of level is again violated, and so tokens of the same kind of component turn out to be at different levels.

Perhaps concerns like these prompted Craver to propose a sufficient condition for sameness in level: "If two things are not related as part to whole, they are not at different levels and so, if they are in the same mechanism, they are, in this very weak sense, at the same level" (2015: 19). Figure 12.5 illustrates the oddity that this condition produces:

According to the sufficient condition, one mechanism, a P_1, is at a different level than the mechanism of which it is a part, X_3, because it fails the necessary condition for sameness of level: it is component of X_3. However, it also meets the sufficient condition for being at the same level as X_1, because the two are in the same mechanism, S, and are not related to each other as part to whole. Thus, a

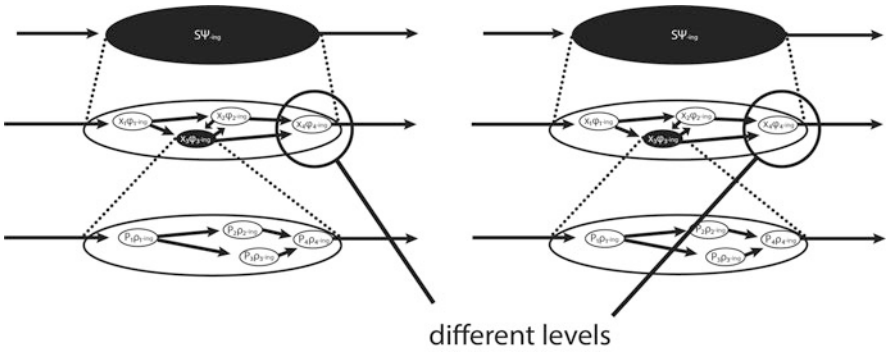

Fig. 12.4 Two identical mechanisms, each an S, with identical components. Craver's necessary condition for sameness of level entails that the two circled components are at different levels

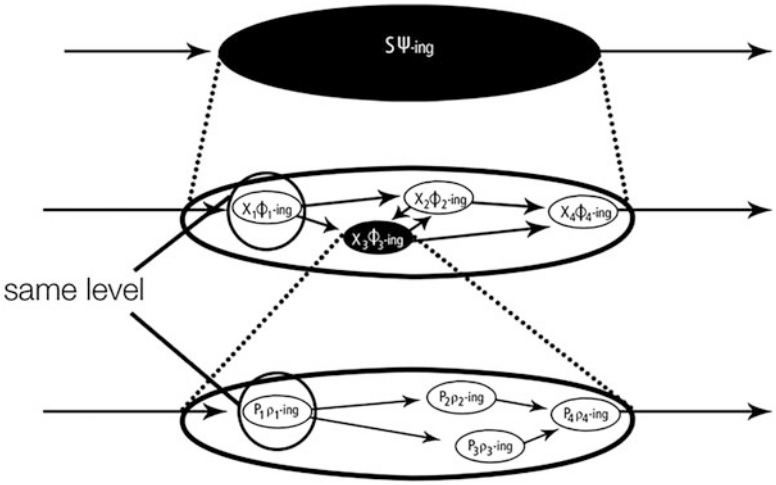

Fig. 12.5 According to the Craver's sufficient condition for sameness of level, a component of one type of mechanism can be at the same level as a mechanism of the same type of which it is a component

token mechanism can be at once at a different level than the token mechanism of which it is a part, and at the same level as a token of the same type of mechanism of which it is a part. Moreover, two tokens of the same type of mechanism (two Xs) can be at the same level, even when one of them, but not the other, is at the same level as something, P_1, that is at a different level from the first. If we regard biceps as mechanisms, then, by Craver's necessary and sufficient conditions for sameness in level, the right bicep and the muscle cell it contains are at different levels, but a token of the same type of muscle cell in the left bicep is at the same level as the right bicep even while also being at the same level as the muscle cell in the right bicep.

Craver is aware of at least some of these peculiarities in his account of levels.[4] However, he sees them not as problems, but as consequences of abandoning levels as Oppenheim and Putnam (and also Wimsatt, 1976) describe them. On the other hand, he also seems not to appreciate some difficulties that his remarks on levels present for other commitments he holds. For instance, he (as well as Craver & Bechtel, 2007), rejects the presence of causal relations between different mechanistic levels, holding that such relations hold only between entities at the same mechanistic level. The reason for this, roughly, is that components do not *cause* the activities of the wholes that they compose: the activities of the components, taken together, *constitute*, rather than cause, the activity of the whole. But, returning to Fig. 12.4, there appears to be no reason why the two circled mechanisms, which are at different mechanistic levels, cannot stand in a causal relation to each other because, consistent with Craver's necessary condition for sameness of level, these mechanisms are at different levels despite not standing in a part-whole relation to each other.

Taken together, Craver's remarks about levels lead to counter-intuitive results – tokens of the same type of thing belonging to different levels, tokens of very different sorts of things belonging to the same level, tokens of the same type of thing being at the same and different levels to a token of another thing – and to inconsistencies – causal relations can and cannot hold between components at different mechanistic levels. Given all this, one must wonder why Craver wishes to insert the idea of levels into his discussion of mechanisms in the first place. If he kept exclusively to a description of mechanisms in terms of parts and wholes, none of the above difficulties would arise. Rather than describing the two identical components in Fig. 12.4 as being at different levels, we say that they are components in different mechanisms. Rather than saying that a token mechanism is at the same and different levels as two tokens of some other type of mechanism, as in Fig. 12.5, we say that it is a component of one and not the other. Rather than saying that causal relations can and cannot hold between components at different mechanistic levels, we say that components cannot cause the activities of the wholes of which they are a part, but they can cause activities in components of other mechanisms. In short, applying levels-talk to an analysis of mechanisms seems only to muddy waters that an appeal to compositional relations alone leave clear.

12.5 Decomposition by Realization

If Oppenheim and Putnam's levels fail to correspond to uniform levels of organization, as Guttman (1976) and Potochnik and McGill (2012) argue, and if they fail to align neatly with discrete scientific branches, as Craver (2007) observes, and if talk of levels seems at best dispensable and at worse harmful in the context of understanding nature's mechanisms, perhaps the *level* concept should be jettisoned.

[4] In a footnote, he mentions that Lindley Darden had raised issues like those I discussed above.

But, if so, does this mean as well that the unity of science hypothesis must be abandoned? What sense, if any, might be made of the unity of science if there exist no distinct levels to be unified? In this section, I introduce the idea of *decomposition by realization* that, in the next section, I'll use as a way to understand what unity of science, absent levels in any of the senses described earlier, might mean. My claim is that the realization relation provides a way to understand how the sciences might be unified, even if, as I suspect, unity of science remains unlikely.

Realization is a relation between a functional description – a description that details some job or activity to be performed – and the object that satisfies that description.[5] For instance, a functional description of a corkscrew is something like: being an entity that can extract corks from bottles. Realization is the relation holding between this description and an object that is capable of extracting corks from bottles. Some philosophers, e.g. Kim (1998b), describe the realization relation as involving orders of properties. The property of being a corkscrew, is the (first-order) property of having some (second-order) property that satisfies the corkscrew description. The first-order property in this case is realized by any one of some, perhaps large, set of second-order properties. Among the second-order realizers of the property of being a corkscrew might be the property of being a waiter's corkscrew, a winged corkscrew, a twist corkscrew, and so on. A functional description, or property, is *multiply* realized when its realizers differ according to some criteria of difference (Polger, 2009; Polger & Shapiro, 2016; Shapiro, 2000, 2004, 2008), as appears to be the case with the realizers of *corkscrew*.

Realizers are not components of what they realize – a waiter's corkscrew is not a component of a corkscrew – and so the realization relation does not lend itself to an analysis in terms of compositional hierarchies in the way that mechanisms do. The realized property is a property of the realizer – being something that extracts corks is a property of a waiter's corkscrew. However, because realizers are themselves often mechanisms, the possibility arises that mechanistic explanations of realizers might yield interesting results. The insight requires a decomposition of realizer-mechanisms in functional terms, where components of the realizer-mechanisms are identified by a functional description, and the functional description is then treated as being realized by some *mechanism* that performs the function. An example will clarify how such a decomposition might proceed.

Grant that the functional description of an eye is *object that collects information about visual properties* (surface reflectances, say). The realizers of VISUAL INFORMATION COLLECTOR are the various mechanisms that perform this function, e.g. camera eyes and compound eyes. Now, focus on one of these realizers, e.g. a camera eye. This eye is a mechanism, consisting of a number of components, each of which can itself be described in terms of the function it performs. Among these components is a LIGHT CONTROLLER that regulates the amount of light that enters the eye. In the camera eye, the realizer of LIGHT CONTROLLER is an iris. The next step in the analysis involves describing the components of the iris in

[5] I use the term 'object' quite generally, to mean property, type, token, event, or whatever.

Functional Descriptions	Realizing Mechanisms
Visual Information Collector	Camera Eye
Light Controller	Iris
Contractors and Expanders	Circular and Radial Muscles
Motion Inducers	Actin-Myosin Filaments

Fig. 12.6 A partial functional decomposition of an eye is on the left side of the table; a decomposition in terms of realizing mechanisms is on the right

functional terms. Two such components are CONTRACTORS and EXPANDERS, which close and open the iris. In the iris, these components are realized, respectively, by circular and radial muscles. In turn, circular and radial muscles have components whose job is to INDUCE MOTION in the muscle, and realizing this job are actin-myosin filaments.

The above analysis of the eye decomposes it in terms of functionally-defined components, and the mechanisms that realize these components; and each such mechanism in turn is then decomposed into functionally-defined components and the corresponding mechanisms that realize them, as in Fig. 12.6:

Crucially, identifying the rows in Fig. 12.6 with levels would be a mistake. The rows in each column are simply related as parts to wholes: a visual information collector contains a light controller; a human camera eye contains an iris. Any effort to force these facts about parts and wholes into the levels format would seem, as it did in the case of Craver's efforts, to invite trouble. Criteria would be necessary to justify why, say, contractors and expanders are at a different level than light controllers, or why circular and radial muscles are at a lower level than the iris. But what these might be, other than those vexed criteria that Craver proposed, are anyone's guess. And what benefit would characterizing the rows in terms of levels bring in addition to what we already have in hand from seeing them as related compositionally?

However, even if we resist the temptation to impose levels on decompositions like those illustrated in Fig. 12.6, we needn't give up on the idea that such decompositions might speak to the prospect of unifying science.

12.6 Realization and the Unity of Science

Whatever else 'unified science' might mean, it must entail a reduction of some sort. A group of allies offers a *unified* front when they act as a single force. Various law enforcement agencies offer a *unified* response to, perhaps, a crime wave when they work as a single team. Victor Emmanuel II *unified* the multiple states in the

Italian peninsula and Sicily in 1861, creating a single country. Unifying science, for Oppenheim and Putnam, meant reducing the number of scientific languages – the proprietary sets of predicates that each science brings to the table when seeking to describe nature – as well as reducing the number of laws necessary for the explanation of natural phenomena. A completely unified science would speak a single language, with its predicates drawn from the lowest-level science, and explain by means of a single class of laws – those that govern the behavior of objects at nature's lowest level of organization.

The criticisms noted in Sect. 12.3 take aim at the levels-model of nature, and the accompanying hierarchy of sciences, that Oppenheim and Putnam thought were required to make sense of the unity of science hypothesis. However, the kind of unity mentioned above, involving language and explanation, does not, *prima facie*, require either of these things. At any rate, it might be possible to take steps toward a *more* unified science without having to endorse the idea of levels. Below I sketch how this can work.

Let's return to the functional decomposition of realizing mechanisms from the preceding section. One realizer of the functional description VISUAL INFORMATION COLLECTOR is a camera eye. Of course, there are many other kinds of realizers of this functional description (depending, to be sure, on the criteria one uses to individuate differences): compound eyes, mirror eyes, corneal refraction eyes, and so on. But suppose, contrary to fact, that only one kind of realizer of this functional description were physically possible. That is, suppose that the only kind of thing that could play the role of a visual information collector were a camera eye. And then imagine further that the functional decomposition of the camera eye into its functional components was similarly constrained, so that irises were the only possible realization of LIGHT CONTROLLER, and circular and radial muscles were the only possible realizers of CONTRACTORS and EXPANDERS. Schematically, the emerging picture looks like Fig. 12.7:

Though similar in appearance to Fig. 12.3, the idea that Fig. 12.7 illustrates is very different. The oval containing the various F(Is) is a realizer of Function Φ, it is not a collection of parts of Function Φ. If Function Φ is VISUAL INFORMATION COLLECTOR, then the oval represents a camera eye, which realizes VISUAL INFORMATION COLLECTOR. However, a camera eye is itself a mechanism, and it can be functionally decomposed into parts like LIGHT CONTROLLER, FOCUSER, IMAGE CONVERTER, each of which corresponds to one of the F(Is). In turn, each of these functional descriptions is, we're imagining, realized by only one kind of mechanism. The LIGHT CONTROLLER, which might be F(x) in Fig. 12.7, is realized by an iris. The FOCUSER, F(w) suppose, is realized by a cornea, although this is not depicted in Fig. 12.7. And then, were Fig. 12.7 to continue to iterate, another series of ovals would appear below the functional decomposition of the iris, each representing the realizers of its functional components, CONTRACTORS and EXPANDERS.

Figure 12.7 stands in contrast to Fig. 12.8, which depicts a situation in which, rather than functional descriptions having single realizers, they have two.

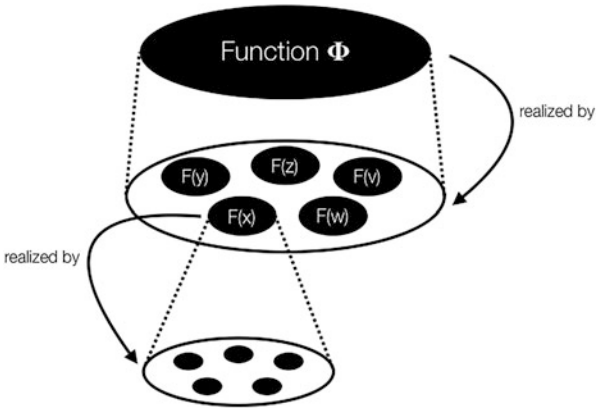

Fig. 12.7 A functional decomposition, in which only one kind of realizer exists for each functional description. In a more detailed diagram, there would be an oval beneath each of the F(Is)

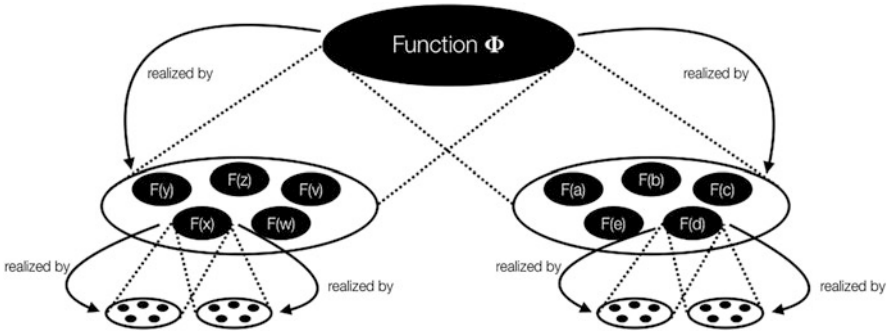

Fig. 12.8 A functional decomposition, in which two kinds of realizer exist for each functional description. In a more detailed diagram, there would be two ovals beneath each of the F(Is)

Assigning once more to Function Φ the description VISUAL INFORMATION COLLECTOR, Fig. 12.8 displays two possible realizers: a camera eye and, suppose, a compound eye. The functional decomposition of the mechanism that realizes the compound eye will of course differ from that of the camera eye. It will include functionally-defined components such as a FOCUSER, and a LIGHT ABSORBER, and a LIGHT GUIDE, which will in turn be realized by mechanisms, some of which may be present in a camera eye, but some not (a cornea, pigment cells, and a rhabdom).

The contrast between Figs. 12.7 and 12.8 provides a way to understand unification that, despite not retaining Oppenheim and Putnam's commitment to levels, promises the sort of reduction in languages and explanations that they took

unification to entail.[6] If there should be only single kinds of realizers that correspond to the elements in a functional decomposition of a VISUAL INFORMATION COLLECTOR, then the science of visual information collectors – of eyes – need speak only of camera eyes and the mechanisms that realize the functional components of camera eyes: irises, circular and radial muscles, actin-myosin filaments. Similarly, understanding how eyes work will require appeal only to those laws, models, and theories that elucidate the behavior of irises, circular and radial muscles, etc. On the other hand, as the number of ways to realize a VISUAL INFORMATION COLLECTOR increases, and the number of ways to realize the functional components of these realizers increases, ever more predicates will be necessary to identify and describe the features of these realizers, and ever more laws, models, and theories will be required to explain their operations. As realization becomes more *multiple*, science becomes less unified.

Of course, it's one thing to articulate a way of unifying science that does not depend on an apparatus of levels, and another to defend the plausibility of the unity of science on this alternative conception. The plausibility depends on answers to at least two questions: (1) How likely is it that the functionally-defined mechanisms that scientists study are singularly, rather than multiply realized? (2) If not singularly realized, how extensively are they multiply realized? Rather than trying to answer these questions, I would like to close this paper reflecting on their significance even should the answers turn out to be: (1) not very likely; (2) very extensively.

For a very long time, at least since Fodor (1974), reductionism has been construed as the thesis that "higher-level" kinds are identical with "lower-level" kinds. The metaphysical possibility of multiple realization is then introduced as a reason to deny such identities, and so ends any hope for reductionism. However, even if the only sensible theory of reductionism entails identities between kinds, one might still wonder why, as Fodor apparently believed, unity of science requires reductionism. On the alternative I have offered, unity of science requires only that the laws governing the world so constrain its contents that, as it happens, functionally-individuated kinds can be realized in only very particular ways.

But, can this contingent fact, if true, be enough to unify science? What of the metaphysical possibility of multiple realization? Even if there existed only one physically possible way to realize a VISUAL INFORMATION COLLECTOR, there must surely be an infinite number of metaphysically possible ways. However, this objection simply returns us to the connection, which I reject, between reductionism and the unity of science. To unify science, one shouldn't have to demonstrate identities between kinds, where these identity claims are vulnerable to metaphysically possible violations. Why should we resist the conclusion that science tends toward unification if, as would be the case should realization be strictly limited, scientists could describe and explain the world with a smaller set of predicates and a narrower set of laws, models, and theories than if multiple realization were the norm? It's preposterous to suggest that metaphysically possible realizers, requiring

[6] Some of the following ideas got their start in Shapiro and Polger (2012).

for their description predicates that don't apply to anything in the world, and for their explanation laws, models, and theories that don't cover anything in the world, should stand in the way of unified science.

References

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Routledge.

Brooks, D. (2017). In defense of levels: Layer cakes and guilt by association. *Biological Theory, 12*(3), 142–156. https://doi.org/10.1007/s13752-017-0272-8

Craver, C. (2007). *Explaining the brain: Mechanisms and the mosaic Unity of neuroscience*. Oxford University Press.

Craver, C. (2015). Levels. In T. Metzinger & J. Windt (Eds.), *Open MIND* (pp. 1–26). https://doi.org/10.15502/9783958570498

Craver, C., & Bechtel, W. (2007). Top-down causation without top-down causes. *Biology and Philosophy, 22*(4), 547–563. https://doi.org/10.1007/s10539-006-9028-8

Eronen, M. (2015). Levels of organization: A deflationary account. *Biology and Philosophy, 30*(1), 39–58. https://doi.org/10.1007/s10539-014-9461-z

Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*(2), 97–115. https://doi.org/10.1007/BF00485230

Glennan, S. (2017). *The new mechanical philosophy*. Oxford University Press.

Guttman, B. (1976). Is 'levels of organization' a useful concept? *Bioscience, 26*(2), 112–113. https://doi.org/10.2307/1297326

Kim, J. (1998a). *Mind in a physical world*. MIT Press.

Kim, J. (1998b). *Philosophy of mind*. Westview Press.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 57*(1), 1–25. https://doi.org/10.1086/392759

Nagel, E. (1961). *The structure of science*. Harcourt Brace.

Oppenheim, P., & Putnam, H. (1958). Unity of science as a working hypothesis. In H. Feigl, M. Scriven, & G. Maxwell (Eds.), *Concepts, theories and the mind-body problem, Minnesota studies in the philosophy of science II* (pp. 3–36). University of Minnesota Press.

Polger, T. (2009). Evaluating the evidence for multiple realization. *Synthese, 167*(3), 457–472. https://doi.org/10.1007/s11229-008-9386-7

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford University Press.

Potochnik, A., & McGill, B. (2012). The limitations of hierarchical organization. *Philosophy of Science, 79*(1), 120–140. https://doi.org/10.1086/663237

Shapiro, L. (2000). Multiple realizations. *Journal of Philosophy, 97*(12), 635–654. https://doi.org/10.2307/2678460

Shapiro, L. (2004). *The mind incarnate*. MIT Press.

Shapiro, L. (2008). How to test for multiple realization. *Philosophy of Science, 75*(5), 514–525. https://doi.org/10.1086/594503

Shapiro, L. (2018). Reduction redux. *Studies in History and Philosophy of Science, 68*, 10–19. https://doi.org/10.1016/j.shpsa.2017.11.004

Shapiro, L., & Polger, T. (2012). Identity, variability, and multiple realizability. In S. Gozzano & C. Hill (Eds.), *The mental and the physical: New perspectives on type identity* (pp. 264–287). Cambridge University Press.

Wimsatt, W. (1976). Reductionism, levels of organization, and the mind-body problem. In G. Globus, I. Savodnik, & G. Maxwell (Eds.), *Consciousness and the brain* (pp. 199–267). Plenum Press.

Chapter 13
Parsimony Arguments in Science and Metaphysics, and Their Connection with Unification, Fundamentality, and Epistemological Holism

Elliott Sober

Abstract Scientists appeal to the principle of parsimony in evaluating competing scientific theories. Philosophers sometimes appeal to a principle by the same name in evaluating competing metaphysical theories. Do justifications for the scientific evaluations carry over to the metaphysical? Here, I outline three "paradigms" for justifying the epistemic relevance of parsimony in science, and assess how they bear on theory evaluation in metaphysics. I connect these issues to questions concerning fundamentality, unification, and Quinean epistemological holism. The upshot is that justifications of parsimony arguments in science often do not apply to the parsimony arguments deployed by metaphysicians. This is not to cast doubt on those metaphysical arguments, but to suggest that they must be made to stand on their own two feet.

Philosophers sometimes use the principle of parsimony to evaluate competing metaphysical theories, and justify their use of the principle by claiming that it is frequently used in science. In doing so, they embrace an instance of the following general thesis:

> **Methodological Naturalism for Philosophy** (MN_p): In evaluating philosophical theories, philosophers should use only the criteria that scientists use when they evaluate theories in the natural sciences (Sober, 2009b, 2015).

This version of naturalism is different from two others:

> **Metaphysical Naturalism**: There are no supernatural entities.
>
> **Methodological Naturalism for Science** (MN_s): Scientific theories should not postulate supernatural entities.

E. Sober (✉)
University of Wisconsin, Madison, WI, USA
e-mail: ersober@wisc.edu

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_13

MN_s gives advice to scientists and makes no reference to how philosophers should go about their business, whereas MN_p gives advice to philosophers and does so by adopting only those norms that are appropriate in science, leaving open whether MN_s is correct.

MN_p seems obviously right in what it says about logically consistency. Scientific theories should be self-consistent, and the same goes for philosophical theories. Should the application of MN_p be more controversial when it comes to the use of a principle of parsimony? Contemporary metaphysicians who have discussed parsimony often think that parsimony comes to the same thing in science and philosophy, so if it is epistemically relevant in the one area it must be relevant in the other (Brenner, 2017; Cowling, 2013; Sider, 2011, 2013).[1] Metaphysicians who take this line have generally not tried to determine in any detail when and why and how the principle of parsimony should be used in science. Indeed, sometimes they maintain that the principle is rock bottom – that it can't be justified in terms of considerations that are more fundamental. Their view seems to be that the principle of parsimony is an article of faith in science, and metaphysicians can therefore embrace that commitment in their own work without fear of embarrassment.

In this paper I'll describe three contexts in which parsimony is *not* rock bottom in science. In each, there are justifications for the claim that simpler theories are better than theories that are more complex in a sense of "better" that is epistemic and not merely pragmatic or aesthetic. I'll then use these three "parsimony paradigms" to investigate whether parsimony in metaphysics can be epistemically relevant in anything like the way it is relevant in science. I'll argue that there is often no carry-over.[2] This is not to cast doubt on parsimony arguments in metaphysics, but to suggest that they must be made to stand on their own two feet. Along the way, I'll discuss metaphysical fundamentality, the virtue of theoretical unification, and epistemological holism.

13.1 Three Parsimony Paradigms

For starters, let's distinguish two formulations of Ockham's razor:

> **Razor of Silence**: If you don't need to postulate that proposition p is true in your explanations, remain silent about whether p is true.

[1] Metaphysicians who reason this way are not thereby committed to MN_p.

[2] Huemer (2009) undertakes a similar inquiry. He describes three ways of justifying the epistemic relevance of parsimony in science; they differ from the ones I'll consider here. I'll explain in what follows why I think the justifications Huemer describes are unsatisfactory. Huemer argues that none of his three vindicates appeals to parsimony that have been used to defend two metaphysical positions – physicalism and nominalism. I disagree with Huemer's bottom line in connection with the mind/body problem (Sect. 13.2.2), but I agree with Huemer's negative verdict about parsimony's relevance to nominalism (Sober, 2009b, 2015). Kriegel (2013), French (2014), Willard (2014), and Thomasson (2015) also doubt parsimony's relevance to metaphysics.

> **Razor of Denial**: If you don't need to postulate that proposition p is true in your explanations, deny that p is true.

The difference between these two razors is as plain as the difference between agnosticism and atheism, but two similarities are worth noting. The first is that they focus exclusively on the task of explanation. Perhaps there are other tasks to which parsimony considerations are epistemically relevant; we'll see one shortly. The second similarity is that both pertain to decisions about what to believe, where believing a proposition is an on/off (dichotomous) state. If degrees of belief are better to use in epistemology than dichotomous belief (as Bayesians usually maintain), then both these razors should be regarded with reserve.

Even if dichotomous belief is an acceptable concept, these razors fail to apply to some of the good parsimony arguments deployed in science. Some of those good arguments concern whether your evidence favors one hypothesis over another, where favoring does not mean that you should believe the one and disbelieve the other. And there's another use of parsimony in science that figures in estimating which of several competing scientific models will be more predictively accurate and which will be less. The wrinkle relevant here is that a model known to be false will sometimes have a higher estimated predictive accuracy than a model known to be true.

The contrast between the razors of denial and silence suggests a *caveat*. One thing to watch out for, in both science and philosophy, are arguments that deploy the razor of denial when the premises seem to license no more than silence. For example, Field (1980) argues that we can dispense with claims affirming the existence of numbers in physics, and draws the conclusion that we should deny that numbers exist. He is a fictionalist about number talk, in keeping with his endorsement of Metaphysical Naturalism. The details and generality of Field's dispensability argument are worth attending to (for example, see Malament, 1982), but there is a separate question – why does dispensability license denial rather than silence?[3]

13.1.1 Paradigm #1 – Parsimony and the Probabilities of Hypotheses

Notwithstanding the fact that the razor of silence uses the dichotomous concept of belief, it, or something like it, has a simple and convincing rationale. If one theory says that proposition A is true, while the other says that the conjunction A&B is

[3] The question of how dispensability is related to the choice between silence and denial is related to the question of when absence of evidence is evidence of absence, on which see Sober (2009a).

true, the first is simpler, and the axioms of probability tell you that

$$\text{For any proposition X, } \Pr(A \mid X) \geq \Pr(A\&B \mid X) \text{, as long as } \Pr(X) > 0.$$

If $\Pr(X) = 0$, the conditional probabilities are not defined. If $\Pr(B|A\&X) < 1$, the first inequality is strict. This means that if you have strong evidence that A is true, and no evidence whatever that bears on B, you should consider believing A rather than the conjunction A&B.

What if A entails B? Then $\Pr(A \mid X) = \Pr(A\&B \mid X)$ as long as $\Pr(X) > 0$. Proposition A may look to be simpler than proposition A&B, and it's true that "A" is syntactically simpler than "A&B" in the language spoken in this paper. However, it would be absurd to say that the first is true and the second is false, or that the first is more probable than the second. Writing "A" rather than "A&B" saves ink, but the commitments of the two are the same. If A entails B, proposition A does not "remain silent" about proposition B.

This is all very neat and tidy, but it is rather useless. Most[4] applications of Ockham's razor involve comparing incompatible hypotheses, and the justification just offered for the razor of silence provides no advice concerning how such problems should be addressed.

There is a second context in which the probabilities of hypotheses are relevant to judging the relevance of parsimony, but this time the hypotheses are incompatible. Medical diagnosticians often embrace the slogan "When you hear hoofbeats behind you, don't expect to see a zebra" (Sotos 1991). This slogan originated in North America where zebras are rare and horses are common. The idea is that if you have two competing diagnoses that both account for a patient's symptoms, and one of them postulates that your patient has a common disease while the other postulates that she has a rare one, the former is better. A simple justification for this principle can be obtained by using the odds formulation of Bayes's theorem, which says:

$$\frac{\Pr(H_1 \mid E)}{\Pr(H_2 \mid E)} = \frac{\Pr(E \mid H_1)}{\Pr(E \mid H_2)} \times \frac{\Pr(H_1)}{\Pr(H_2)}.$$

This equation says that the ratio of the posterior probabilities of the two hypotheses equals the likelihood ratio times the ratio of prior probabilities. If all you know about your patient is that she belongs to a population in which disease C is common and disease R is rare,[5] and you are prepared to view her as a random sample from that population, this justifies your assigning the hypothesis that your patient has C a much higher prior probability than the prior probability you assign to her having R. If the two diagnoses fit the observed symptoms equally well (so their likelihoods

[4] I'll explain why I say "most" rather than "all" in Sect. 13.1.3.

[5] If you know that your patient belongs to multiple populations, "the problem of the reference class" arises. I ignore that problem here, as it doesn't bear on the point I'm making.

are equal), then the posterior probability that your patient has C will be much larger than the posterior probability that your patient has R.

That said, you still might want to question whether hypothesis C really is *simpler* than hypothesis R. Return to the analogy: it's true that horses are more *common* than zebras (at least in North America), but that doesn't mean that horses are *simpler* than zebras. I agree, but this point probably won't stop physicians and others from thinking that the horses-rather-than-zebras principle is an instance of Ockham's razor, so it is well to list it here and recognize that it makes sense. Notice I've justified the principle by linking probability assignments to frequency data. The reason you should assign C a higher prior probability than you assign to R isn't *a priori*. Notice also that the conclusion of the probability argument is not that H_1 is true and H_2 is false, nor that you should believe H_1 and disbelieve H_2; rather, the conclusion is that H_1 is far more probable than H_2, conditional on the observed symptoms, which the two hypotheses explain equally well.

The two examples discussed so far, in which the simpler of two hypotheses has the higher posterior probability, are modest, but the demonstrations of their epistemic relevance are, I think, compelling. Alas, the picture switches from rabbit to duck when you turn to more ambitious attempts to show that simpler theories have higher prior probabilities. Harold Jeffreys' simplicity postulate (see, for example, Wrinch & Jeffreys, 1921) is the most famous and influential attempt to implement this idea. He proposed a simplicity ordering of an infinite set of mutually incompatible mathematical equations, assigning higher prior probabilities to simpler hypotheses and lower priors to hypotheses that are more complex, making sure that the sum of these infinitely many probabilities equals 1. Jeffreys' proposal is impressive in its ambitions, but it is entirely unconvincing. Why this assignment of priors rather than other ones?[6] I believe that this question has never been answered satisfactorily.

I've listed parsimony in relation to the probabilities of hypotheses as the first parsimony paradigm, but I think it plays third fiddle to the other two.

13.1.2 Paradigm #2 – Parsimony and the Likelihoods of Hypotheses

The second paradigm involves comparing the likelihoods of simpler and more complex hypotheses. One of the main examples that interests me is the comparison

[6] Huemer (2009, p. 220) describes a proposal that is in the spirit of Jeffreys. He points out that if there is a simplest theory in an infinite list, but no upper bound on how complex a theory can be, then as you move from simpler to more complex, the probabilities assignments must "generally" decline. But why should the decline be *strictly* monotonic? Huemer says that it would be "arbitrary" to assign the maximum prior probability to one of two theories, neither of which is maximally simple, whereas it would be "natural" to assign the largest prior probability to a theory that is maximally simple. I think that both decisions are arbitrary.

of a common cause hypothesis with a hypothesis that postulates separate causes. The former is more parsimonious because it postulates one cause rather than two (or more). Consider for example, the hypothesis that human beings and chimpanzees have a common ancestor (CA) in comparison with the hypothesis of separate ancestry (SA) — that they have no common ancestor. The data that moves biologists to favor the first hypothesis over the second concern the similarities that the two species share. Here I use "favors" in the technical sense deployed by the Law of Likelihood (Hacking 1965; Sober, 1988, 2015). Applied to the example at hand, the law says:

> (LoL) The fact that human beings and chimpanzees both have trait T favors the hypothesis of common ancestry (CA) over the hypothesis of separate ancestry (SA) precisely when Pr(human beings and chimpanzees have trait T | CA) > Pr(human beings and chimpanzees have trait T | SA).

What matters here is not how probable the hypotheses are, given the observation, but how probable each hypothesis says the observation is. This is what the technical term "likelihood" means in statistics – a confusing choice of terminology if ever there was one.

Assumptions inspired by Reichenbach (1956), which are plausible in this evolutionary context, entail that the likelihood inequality in LoL is correct (Sober, 1988; Sober, 2015). To state these assumptions, I'll use the letters x, a, y, b, and c shown in Fig. 13.1; they represent probabilities. Focusing for the moment on the CA hypothesis, I'll explain what those letters mean:

x = Pr(humans have a tail bone | C has a tail bone)
a = Pr(humans have a tail bone | C lacks a tail bone)
y = Pr(chimpanzees have a tail bone | C has a tail bone)

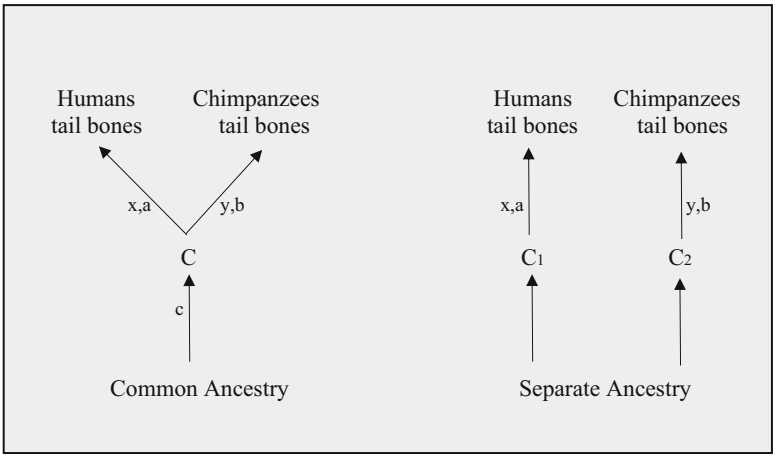

Fig. 13.1 Two Genealogical Hypotheses

b = Pr(chimpanzees have a tail bone | C lacks a tail bone)
c = Pr(C has a tail bone).

Here are the assumptions:

1. Intermediate probabilities: $0 < x,a,y,b,c < 1$.
2. Screening-off: states of descendants are probabilistically independent of each other, conditional on each possible state of their ancestor.
3. Cross-branch homogeneity: (x-a) and (y-a) are nonzero and have the same sign.
4. Cross-model homogeneity: the conditional and unconditional probabilities that a species has a trait are the same in the CA and the SA hypotheses.

Assumption 4 ensures that the common ancestry and the separate ancestry hypotheses are alike, save for the fact that one postulates common ancestry and the other postulates separate. This assumption allows one to isolate the evidential import of parsimony by having the two models differ in parsimony while being the same with respect to their other properties. Given these four assumptions, one can deduce that the more parsimonious hypothesis has the higher likelihood.[7]

Assumptions 1–4 are not *a priori*, and real-world systems can easily violate them, with the result that the observed "matching" of two entities fails to favor the common cause hypothesis over the hypothesis of separate causes. Here are two such examples:

- Suppose there is a gene G the possession of which renders a woman permanently sterile after she becomes pregnant. Suppose you observe that two women have gene G. This observation favors the hypothesis that they are not sisters over the hypothesis that they are.
- Suppose that there are 100 businesses in Madison, where each has one boss and several employees. You observe that two employees in this ensemble, Jack and Jill, are unhappy with their bosses. This similarity favors the hypothesis that they have different bosses over the hypothesis that they have the same boss if Jack and Jill differ in the following respect. Jack's probability of being unhappy with his boss increases if his boss is demanding, whereas Jill's probability of being unhappy with her boss declines if her boss is demanding.

These examples highlight the fact that the likelihood argument does not show that any old common cause hypothesis has a higher likelihood than any old separate cause hypothesis. Unfortunately, many intuitively plausible published examples that are said to illustrate how common cause hypotheses have higher likelihoods fail to supply enough details to secure the desired conclusion.

There are two broader lessons that emerge from the likelihood analysis of common cause and separate cause hypotheses. The first addresses an idea that

[7] All traits shared by human beings and chimpanzees are evidence favoring CA over SA if they satisfy the assumptions just mentioned, but they may differ from each other with respect to how strongly they favor CA over SA. For a likelihood account of how degree of favoring works in this case, see Sober and Steel (2017).

is fairly standard in philosophy of science – that fitting the data and parsimony are separate and independent considerations relevant to hypothesis evaluation. The point is not correct in the present context, where fit-to-data is standardly understood in terms of likelihood, and differences in parsimony are mirrored in differences in likelihood. The idea that parsimony is a "super-empirical virtue" is wrong when the comparison is between common-cause and separate-cause explanations.[8]

The second lesson concerns assumption (1), that all five probabilities are intermediate. Violate that assumption and it's easy to find cases in which the likelihoods of CA and SA are equal. This result goes contrary to the following idea: if parsimony is evidentially relevant in the case of probabilistic hypotheses, it also must be relevant in the case of deterministic hypotheses. This idea may be intuitive, but the intuition is wrong. There is no automatic carry-over; the shift from nonextreme probabilities to extreme probabilities makes a profound difference. This simple point provides a useful warning, to which I will return.

13.1.3 Paradigm #3: Parsimony and the Predictive Accuracies of Models

Scientists often find that a complex model easily accommodates the data at hand, but then turns out to do a very poor job of predicting new data drawn from the same underlying reality. When this happens, scientists say that the complex model "overfit" the old data. This lived experience has led scientists and statisticians to realize that finding models that are predictively accurate requires a balancing act, one that takes account of both fit-to-data and parsimony. But how is the "right" balance to be struck? Is the decision an arbitrary matter of taste? Statisticians have developed a variety of techniques for coping with this problem. There are mathematical explanations for why fit-to-data and parsimony trade off against each other. And given some assumptions, one can obtain mathematical proofs concerning what the optimal trade-off ought to be.

Here's a simple example that illustrates the kind of prediction problem I have in mind. You are driving through farmland south of Madison and stop your car when you see that there are two vast fields of corn on either side of the road. You choose 100 corn plants from each field and measure their heights. You see that the average height in your sample from the first field is 62 inches and the average height in your sample from second is 66 inches, so the difference in height between the

[8] Philosophers often confuse the fact that a hypothesis fits the data with its being logically consistent with the data. Fit is a matter of degree whereas logical consistency is not, but more importantly, the fact that hypothesis and observation are logically consistent with each other is perfectly compatible with the observation's being evidentially irrelevant, and also with its disconfirming. Two hypotheses that each entail the data fit it equally well.

two samples is 4 inches. You want to use these observations to predict what you'll observe if you draw a second sample of the same size from each field.

To make your task more concrete, you decide to consider two models of how the average heights (h_1 and h_2) in the two fields are related to each other:

(NULL) $h_1 - h_2 = 0$
(DIFF) There exists a number d such that $h_1 - h_2 = d$.[9]

NULL is properly so called, since it says that there is no difference between the average heights in the two fields. My name for the second model is a bit misleading, however, in that DIFF doesn't say that the average heights differ; it says that they *can* differ, but they need not. DIFF is a near tautology, but as we'll now see, that doesn't mean that it makes no predictions. Similarly, Null is almost certainly false, but that doesn't mean that it will make inaccurate predictions.

It is clear what Null predicts about the new observations you are going to make, but what does DIFF predict? To use DIFF to make a prediction, scientists typically estimate the value of parameter d by looking at the data at hand; they then plug that estimate into DIFF to predict the new data you'll soon obtain. But what estimation procedure should you use? A standard practice is to use the method of maximum likelihood estimation. That is, one wants to find the estimate that maximizes the probability of what you've observed. The maximum likelihood estimate, given the data you have, is that $d = 4$ inches.

DIFF contains an adjustable parameter, but when you fit the model to your data, you obtain the fitted model f(DIFF), which contains no adjustable parameter. DIFF, by itself, predicts nothing, but with help from the data at hand, you can construct f(DIFF). Notice that f(DIFF) fits the data perfectly, while NULL does not. Indeed, DIFF can perfectly accommodate the data, no matter what the observations are. Notice in addition that DIFF is more complex than NULL if you measure complexity by counting adjustable parameters. NULL has zero while DIFF has one. So NULL has one point in its favor and one against; ditto for DIFF.

H. Akaike (1973) proved a remarkable theorem about predictive accuracy:

> Akaike's Theorem: An unbiased estimate of the predictive accuracy of model $M = \log(\Pr(\text{data} \mid f(M))) - k$.

What does "unbiased" mean? When an unbiased estimator is repeatedly applied to different data sets drawn from the same underlying reality, the estimates obtained will tend to be "centered" on the true value. A kitchen scale is unbiased if repeatedly weighing an apple that weighs 6 ounces will, on average, output the estimate that the apple weighs 6 ounces. The "log" in Akaike's theorem denotes the natural

[9] Notice that NULL entails DIFF. This isn't crazy; scientists often test nested models against each other. I'll explain the nonBayesian rationale for doing so shortly. However, the comparison of NULL and DIFF by using the model selection criterion I'll describe would be unchanged if NULL were compared with DIFF *, where DIFF* stipulates that $d \neq 0$.

logarithm, and "k" represents the number of parameters in the model. The minus sign in Akaike's theorem indicates that models are penalized for their complexity.

Akaike's theorem, like any theorem, is derived from assumptions. Here are the ones that Akaike used in his proof:

- M is true.
- Repeated samples are drawn from a single underlying reality.
- A "normality" assumption – that repeated estimates of a parameter in a model will form a bell-shaped distribution.

The first of these assumptions can be weakened if one is interested merely in ordering the predictive accuracies of two or more competing models. It can be replaced with the assumption that one of the models is close to the truth. Nodding to Hume, Forster and Sober (1994) called the second assumption a "uniformity of nature assumption."

Akaike's theorem led to the formulation and use of a procedure for estimating the predictive accuracies of competing models:

The Akaike Information Criterion (AIC): $\text{AIC}(\text{M}, \text{data}) = \log(\Pr(\text{data} \mid \text{f}(\text{M}))) - \text{k}$.

Since AIC says that there are two considerations relevant to comparing models for their predictive accuracies, it is not inevitable that the more parsimonious model receives the better AIC score. That depends on the data. In our example, f(DIFF) has a higher log-likelihood than NULL. The question is whether the difference in the log-likelihoods of the two models is *sufficiently* large to tip the scales in favor of the more complex model. More specifically, the question here is whether the difference in the two log-likelihoods is greater than 1.

It's important to recognize that Akaike's theorem does not entail that AIC is the best criterion for estimating predictive accuracy. There are other properties one might want an estimator to have besides unbiasedness. For example, an unbiased estimator may have a large variance or a small one; variance represents the amount of dispersion from the mean value that repeated estimates will exhibit. A good kitchen scale should have small variance. So the theorem doesn't suffice to justify AIC. What is more, it is arguable that biased estimators aren't always beyond the pale; if they are not, then good estimators may not need to be unbiased.[10]

These ideas from model selection theory (which might better be called "the theory of model evaluation") have an interesting application to the question of why postulating one cause for a given effect is better than postulating two. In the previous section on likelihoods, I discussed cases in which there are two (or more) observations and the competition is between a common cause hypothesis and the hypothesis of separate causes. The problem I'll discuss now involves a single effect, and one wants to compare a hypothesis that postulates one cause with a hypothesis that postulates two.

[10] See Vassend, Sober, and Fitelson (2017) for discussion of the trade-off between bias and variance in statistical decision theory.

Table 13.1 Pr(the subject has lung cancer | —)

	The subject inhaled asbestos.	The subject did not inhale asbestos.
The subject smoked cigarettes.	$x + c + a$	$x + c$
The subject did not smoke cigarettes.	$x + a$	x

Table 13.2 Four causal models concerning how cigarettes and asbestos affect lung cancer

Models		Number of adjustable parameters
BOTH	a = α and c = σ	2
ASBESTOS-ONLY	a = α and c = 0	1
CIGARETTES-ONLY	a = 0 and c = σ	1
NULL	a = 0 and c = 0	0

How should the following four hypotheses about the causal impact of smoking cigarettes and inhaling asbestos on lung cancer be evaluated?

BOTH: Smoking and Asbestos both cause lung cancer.
SMOKING-ONLY: Smoking causes lung cancer, but asbestos does not.
ASBESTOS-ONLY: Asbestos causes lung cancer, but smoking does not.
NULL: Neither smoking nor asbestos causes lung cancer.

Each hypothesis makes a claim about the relationships of the probabilities represented in Table 13.1. The symbol "x" denotes the baseline probability – the probability of getting lung cancer if you don't smoke cigarettes or inhale asbestos. The difference made by cigarettes is represented by "c"; the difference made by asbestos is represented by "a". For simplicity, I'll assume additivity, meaning that the difference made by smoking and asbestos together is just the sum of the differences that each would make on its own.

The notation introduced in Table 13.1 allows you to formulate the four causal models shown in Table 13.2. Models that say that asbestos makes no difference assert that a = 0, whereas models that say that asbestos might make a difference say that a = α, where α is an adjustable parameter. Similar remarks pertain to what models say about cigarettes; they'll either say that c = 0 or say that c = σ, where σ is an adjustable parameter.[11] Notice how the models differ with respect to their number of adjustable parameters (Forster & Sober, 1994; Sober, 2015).

AIC can be used to evaluate these four models. What is required is a data set in which the observed frequency of lung cancer is recorded for each of four "treatments." These four frequencies need not sum to 1. Maximum likelihood estimates of adjustable parameters can be obtained from this data set. The BOTH model will fit the observations perfectly, no matter what those observations are.

[11] It wouldn't affect the analysis if α and σ were required to be nonzero (or positive) when the model doesn't say that they are zero.

However, that doesn't mean that BOTH will have the best AIC score. After all, BOTH pays a penalty for complexity that is larger than the penalties paid by the other three models. If two of these models fit the data equally well, one of them may be more parsimonious than the other; if so, the more parsimonious model will have the better AIC score. However, AIC is not a mere tie-breaker; it applies to the case in which competing models differ in their goodness-of-fit.

I hope it is clear how parsimony paradigms 2 and 3 differ. In the likelihood paradigm, simpler hypotheses have higher likelihoods; in the AIC paradigm, simpler models do not have higher likelihoods. What has a likelihood in AIC is a *fitted* model, not a model that contains adjustable parameters. Parsimony enters AIC, not by comparing likelihoods, but by comparing the number of adjustable parameters that models contain.[12]

There are several interesting conceptual questions about AIC that I won't delve into here. The present point of importance is that AIC shows how parsimony can be epistemically relevant – not to saying which model is true or probably true, but to estimating a model's predictive accuracy.

[12] Huemer (2009, pp. 221–223) describes a likelihood justification of parsimony where the hypotheses considered are models with adjustable parameters. He says that this justification is "the most promising" of the three he considers. Huemer isn't talking about AIC here; rather, he is suggesting a Bayesian rationale for valuing simpler models. He considers a simple model S and its more complex competitor C, and says that "the likelihood account argues that S *typically* has the higher likelihood P(E|S). Since S is compatible with a smaller range of data, it assigns a higher *average* probability (or probability density) to those possible of data which it allows. C spreads its probability over a larger range of possibilities, consequently assigning a lower probability (density), *on average*, to the possibilities which it allows (italics mine)." However, on the next page Huemer notes that "even when the simpler of two theories fits a narrower range of data than the more complex theory, the simpler theory need not have a higher likelihood in relation to every possible datum which both accommodate: rather, the simpler theory must have a higher *average* likelihood within the range of data which it accommodates than the complex theory has within the range which the complex theory accommodates." Huemer offers no argument for the claim that simpler models "typically" have the higher likelihood, given the data at hand. It is consistent with Huemer's point about "averages" that, in each situation in which S assigns positive probability densities to a smaller range of values than C does, that S has a lower likelihood than C across *almost all* of that value range. The desired result would be obtained if both models were required to impose flat probability density distributions on their parameters, a possibility that Huemer represents in a figure. Bayesians often embrace this stipulation; frequentists demur. Another limitation of Huemer's argument is that he is discussing models in which each parameter is assigned a specified finite value range. This is often not the case, as in his example of LIN and PAR.

13.2 How Philosophical Parsimony Arguments Measure Up – Three Examples

13.2.1 Is Observation the Elephant in the Room?

It may seem that the three parsimony paradigms I've described are nonstarters for metaphysics, since that subject is largely *a priori*. The sticking point is that the likelihood paradigm (1.2) and the model selection paradigm (1.3) both focus on hypotheses that make predictions about *observations*. True, the paradigm described in Sect. 13.1.1, which holds that simpler hypotheses have higher probabilities, does make a little room for *a priori* facts about probability (e.g., that the simpler hypothesis A can't be less probable than the more complex hypothesis A&B), but that doesn't seem to help much.

One part of this problem can be solved by recognizing that the concept of observation need not be narrow. Harman (1977) talks about the example of a gang of hoodlums setting fire to a cat. You see this shocking event and instantly believe that what the hoodlums are doing is wrong. Once we acknowledge that observations are "theory-laden", we can count "what the hoodlums are doing is morally wrong" as an observation statement. All I mean here by "theory-laden" is that to observe that a proposition p is true, you need to understand the concepts used in the proposition. That understanding takes the form of grasping propositions that together count as your "theory."

As this example illustrates, normative theories in ethics (e.g., utilitarianism) can be evaluated by seeing what they predict about observations. If you think that proposition p is true and utilitarianism entails that it is, that's a plus for utilitarianism. And if you think that proposition q is false and utilitarianism entails that q is true, that counts against utilitarianism. Of course, these so-called observations need not be absolutely certain. Some may be mistaken and a good philosophical theory may provide a reason for you to change your mind about what you initially thought was a true observation statement. This is the point about reflective equilibrium that Goodman (1965) and Rawls (1971) made familiar.

There is a different and more challenging problem posed by the two most promising parsimony paradigms. That more pressing concern arises because in both those paradigms, the epistemic relevance of parsimony to the evaluation of competing hypotheses depends on differences in parsimony mirroring differences in how hypotheses are probabilistically related to observations (even when "observation" is construed broadly). This will be a central issue in what follows.

13.2.2 Two Pretty Successful Parsimony Arguments in Philosophy

Although many parsimony arguments in philosophy don't work out very well when judged by the standards satisfied by compelling parsimony arguments in science, some do. A fairly straightforward example of the latter is the evidential argument from evil. It claims that the amount of evil that exists in this world is evidence against the existence of an all-powerful, all-knowing, and all-good (all-PKG) deity. This claim about evidence can be given a likelihood formulation (Draper, 1989; Sober, 2004, 2018):

$$\Pr(\text{E} \mid \text{an all-PKG God exists}) < \Pr(\text{E} \mid \text{no all-PKG God exists}) .$$

Here E is a reasonably detailed description of the kinds and quantities of evils that there are, not the bland proposition that there is some evil in the world. This argument doesn't mention parsimony, but if "an all-PKG God exists" is less parsimonious than "no all-PKG God exists," we have here a case in which the more parsimonious hypothesis has the higher likelihood. In this respect, the evidential argument from evil and arguments for common ancestry in evolutionary biology are on the same page.

A less straightforward example of a philosophical parsimony argument that approximates the good parsimony arguments used in science can be found in the mind-body problem. The mind/brain identity theory, functionalism, and dualism are usually formulated as follows:

(Identity)	Every mental property is identical with some physical theory.
(Functionalism)	Every mental property strongly supervenes on some physical property, and mental properties are multiply realizable.
(Dualism)	Mental properties are not identical with physical properties, nor do mental properties strongly supervene on physical properties.

I doubt that the parsimony differences among these three theories are epistemically relevant, but I do think that *instantiations* of each of these theories differ in parsimony in a way that is. To see why, let's apply these three big-picture theories to the old-fashioned example of pain and c-fiber firing:

(Identity*)	Being in pain is one and the same property as having one's c-fibers fire.
(Functionalism*)	Being in pain strongly supervenes on having one's c-fibers fire, but c-fiber firing is not the only supervenience base that being in pain has.
(Dualism*)	Being in pain is not identical with having one's c-fibers fire, and pain does not strongly supervene on c-fiber firing.

What do these three theories predict you'll observe in a data-gathering study in which you track 100 subjects through an entire year? On each day, you use a phone

Table 13.3 Frequencies obtained in an experiment

	The monitor says that c-fibers are firing.	The monitor says that no c-fibers are firing."
The subject pushes the button that says "pain."	f_1	f_2
The subject pushes the button that says "no pain."	f_3	f_4

Table 13.4 The probabilities of four conjunctions

	The subject's c-fibers are firing	The subject's c-fibers are not firing
The subject is in pain	p_1	p_2
The subject is not in pain	p_3	p_4

app to ask each subject whether they are in pain, and use a monitor that the subjects have agreed to wear that detects whether their c-fibers are firing? For each subject, you therefore obtain 365 pairs of observations, so you obtain 36,500 observation pairs in toto. These allow you to compute the values of the four frequencies ($f_1 \ldots f_4$) represented in Table 13.3. Each cell entry represents how frequently this or that conjunction is true.

The next step is to construct models of the subjects' inner states. These inner states are described in Table 13.4; four conjoint states are described, and each has one of four probabilities ($p_1 \ldots p_4$) that sum to one. The identity theory for pain and c-fiber firing says that two of the four conjoint events cannot occur, whereas dualism doesn't rule out any of them. This means that the identity theory has fewer adjustable parameters:

(I) $p_1 + p_4 = 1$
(D) $p_1 + p_2 + p_3 + p_4 = 1$

Notice that the models differ in their number of adjustable parameters; there is one such parameter in the I model (since fixing the value of p_1 automatically settles what the value is for p_4 in that model), and there are three in D.[13]

How can these two models be connected with the frequency data from your study? If your observations were error-free, the observed frequencies would furnish maximum likelihood estimates of the probabilities in these two models. However, your observations may be subject to error. Here's a simple representation of the possibility of error:

Pr(the subject pushes the "no pain" button | the subject is in pain) $= r_1$
Pr(the subject pushes the "pain" button | the subject is not in pain) $= r_2$
Pr(the brain monitor says "c-fibers are not firing" | c-fibers are firing) $= r_3$
Pr(the brain monitor says "c-fibers are firing" | c-fibers are not firing) $= r_4$

[13] The D model might be constrained to require that $p_1 < 1$, but that won't affect the AIC comparison of the two models.

I'll assume that you have good estimates of these four probabilities, obtained from other studies.

The upshot is that your frequency data from the study you ran provides maximum likelihood estimates of the parameters in the I and D models. The D model will fit the data better than the I model, regardless of what the data look like. However, D has more adjustable parameters than I. AIC takes both facts into account in deciding which model is better.

A functionalist model of the relationship of pain and c-fiber firing takes the following form:

(F) $p_1 + p_2 + p_4 = 1$

The F model says that $p_3 = 0$, since the model claims that pain strongly supervenes on c-fiber firing. This model allows that p_2 may be positive, since multiple realizability means that individuals may be in pain even though they don't have c-fibers.[14] The F model has two adjustable parameters, so it is more parsimonious than the D model but less parsimonious than the I model.[15,16]

Dualism is often said to have no problem with observed correlations of mental and physical states. This is typically viewed as a virtue of dualism, but according to AIC it is a serious flaw. However, the flaw needn't be fatal; whether the D model is the worst of the three, or the best, or is middling, depends on the data.

If the F model for the relationship of pain and c-fiber firing fails to receive the best AIC score, that doesn't mean that Functionalism is false. Maybe pain doesn't supervene on c-fiber firings, but supervenes on some other physical state. Functionalists will want to look for these. The same holds for Identity theorists. The situation for dualism is a bit different. If the D model for pain and c-fiber firings has the worst AIC score, then it doesn't make much sense to look for a physical state other than c-fiber firings that fits what dualism says. Viewed in this way, the identity theory and functionalism are the guiding principles of two research programs (Lakatos, 1978), which aren't automatically undermined when a single model turns out to be flawed. It's hard to see what the dualist research program is; dualism seems to be mired in nay-saying, having nothing positive to contribute.

My conclusion is that parsimony is epistemically relevant to evaluating competing hypotheses about the mind-body relation, but the locus of its relevance is specific models that implement the directives of broad *isms*, not the *isms* themselves. The latter do differ in how parsimonious they are, but the three parsimony paradigms I've described do not show how those parsimony differences are epistemically relevant.

[14] Requiring that $p_2 > 0$ would not affect the AIC comparison of models.

[15] For simplicity, I've treated pain and c-fiber firing as on/off states, but the three mind/body theories can also be applied to claims about the relationship of the *intensity* of pain to the *frequency* of c-fiber firings.

[16] Huemer (2009) discusses the mind/body problem and concludes that the parsimony arguments used there don't correspond to the ones that are applicable in science. He does not consider the use of AIC just described.

It may seem odd to apply AIC to the mind/body problem. After all, metaphysicians aren't interested in predictive accuracy; they aim at truth! I earlier described how AIC fits nicely into an instrumentalist philosophy of science, but AIC also connects with scientific realism, where the goal is to figure out which of a set of candidate theories is closest to the truth. AIC estimates predictive accuracy, but it also estimates a model's Kullback-Leibler distance from the truth; for discussion, see Forster and Sober (1994) and Sober (2015). In the context of pain and c-fiber firings, the question is which model of their relationship is closest to the truth.

13.2.3 A Failed Parsimony Argument Concerning Mental Causation

If c-fiber firing causes wincing, isn't it redundant to additionally assert that pain causes wincing? If so, does the principle of parsimony license the conclusion that you should deny that pain causes wincing? Before you wield the razor, it is important to consider a second question. If pain causes wincing, isn't it redundant to add that c-fiber firing causes wincing? If so, does the principle of parsimony license the conclusion that you should deny that c-fiber firing causes wincing? The principle of parsimony seems to tell you to cut something, but it's unclear what to cut. Maybe the solution is to note that you need to invoke physical causes for lots of events that aren't caused by mental states (like the freezing of a lake in winter), so maybe the best way forward is to deny the causal efficacy of pain.

This dialectic has the outward trappings of a well-motivated scientific parsimony argument. It sounds just like the example discussed in Sect. 13.1.3 concerning asbestos exposure, cigarette smoking, and lung cancer, but the two examples differ in a way that wrecks the parsimony argument against mental causation. To see why, let's consider four causal models:

BOTH: Pain and c-fiber firing cause wincing.
PAIN-ONLY: Pain causes wincing, but c-fiber firing does not.
CFF-ONLY: C-fiber firing causes wincing, but pain does not.
NULL: Neither pain nor c-fiber firing causes wincing.

Each of the causal hypotheses can be associated with a probability model. To state these models, I'll use the terminology depicted in Table 13.5. If a subject's c-fibers are not firing and they are not in pain, the probability that they wince is x; this is the "baseline" probability, which gets used to talk about how the probability of wincing changes if there is c-fiber firing, or pain, or both. For example, p is the difference in the probability of wincing caused by shifting from no pain to pain. The models we'll consider are additive. They are described in Table 13.6.

In the example concerning cigarettes, asbestos, and lung cancer, frequency data allow adjustable parameters in models to be estimated. This is because there are individuals in each of the four treatment cells shown in Table 13.1. Unfortunately,

Table 13.5 Pr(S winces | —)

	S's c-fibers are firing	S's c-fibers are not firing
S is in pain	$x + c + p$	$x + p$
S is not in pain	$x + c$	x

Table 13.6 Four causal models concerning how pain and c-fiber firing affect wincing

Models		Number of adjustable parameters
BOTH	$p = \alpha$ and $c = \beta$	2
PAIN-ONLY	$p = \alpha$ and $c = 0$	1
CFF-ONLY	$p = 0$ and $c = \beta$	1
NULL	$p = 0$ and $c = 0$	0

this isn't possible for the present problem, if pain strongly supervenes on c-fiber firing. In that case, there will be no data on the frequency of wincing among people whose c-fibers are firing though they are not in pain. This means that a probability in the left-hand column of Table 13.5 is not defined, and this has the consequence that the AIC scores for the four models in Table 13.6 are not defined.

The question of whether pain or its supervenience bases cause wincing is ill-formed (Shapiro & Sober, 2007). This point generalizes. If Y supervenes on X, don't conclude that the hypothesis that Y causes Z and the hypothesis that X causes Z are in competition. And don't claim that affirming one hypothesis and denying the other is sanctioned by Ockham's razor – at least not if the razor is understood in terms of the three parsimony paradigms I discussed earlier.[17] This has implications for a perennial question in philosophy of social sciences concerning methodological holism and methodological individualism (Wright, Levine, and Sober 1992), and more generally for the question of how "levels of organization" should be understood in science.

13.3 Parsimony and Fundamentality

Fundamentality is and has been an important question in metaphysics (Sider, 2011, 2013). Competing hypotheses about what is fundamental can of course be evaluated for their parsimony, but when are parsimony differences between such hypotheses epistemically relevant?

I'll use the term "entity" to encompass token objects, as well as properties, relations, and propositions. What does it mean to distinguish entities that are fundamental from ones that are not? The rough idea is that the properties of the former determine the properties of the latter, but not conversely. The nature of the

[17] This doesn't mean that there is no way to test the epiphenomenalist hypothesis that pain is a correlate of wincing, not a cause. For details, see Shapiro and Sober (2007) and Sober (2015).

determination relation will shift somewhat from one class of entities to another. However, in all these cases, fundamentality is *factive*. For example, phlogiston can't be fundamental if there is no such thing, and the principle of sufficient reason can't be fundamental if it is false.

The fundamentality of an object entails that it exists, but the converse does not hold. Here I am speaking English. This leaves room for metaphysicians to argue, for example, that composite objects do not exist, but the argument needs to go beyond the fact that they aren't fundamental. In keeping with this idea, I want to suggest that the existence of multiple "levels of organization," which are routinely recognized in science, is compatible with the general claim that macro objects, properties, and processes strongly supervene on micro objects, properties and processes. What is more, causality can "cross levels." Micro events can cause macro events, and *vice versa*. Scientists frequently make such cross-level claims; some such claims may be false, of course, but there is nothing *a priori* false or incoherent about them. The caveats registered in the previous section in connection with pain and wincing are relevant here.

The idea of fundamentality has an anti-redundancy clause built into it. If X determines Y (in the relevant sense), but not conversely, the claim that both X and Y are fundamental is false. You don't need to invoke the principle of parsimony to draw that conclusion.[18] However, suppose you don't already know whether X determines Y but not conversely. Is this an opening for the principle of parsimony to do some work? Maybe the hypothesis that "X is fundamental and Y is not" is more parsimonious than the hypothesis that "X and Y are both fundamental," but why does that indicate that the former is true and the latter is false?[19]

When scientists confront problems of this sort, they don't reach for their razors. More likely, they endeavor to prove that X determines Y but not conversely, or prove that X fails to determine Y. An instructive example is the centuries-long puzzle of whether the parallel postulate can be derived from the other axioms and postulates in Euclidean geometry. It would be more parsimonious to hold that the parallel postulate is derivable, but that was not how geometers argued. Rather, they repeatedly tried and failed to prove the parallel postulate, and they finally were able to prove that the parallel postulate is independent of the other axioms and postulates. Proof was the gold standard, and waving a razor would have rightly been regarded as a distraction.

[18] This raises a question about the status of "Ockham's Laser," which Schaffer (2015) and Bennett (2017) suggest is the proper parsimony principle in metaphysics – do not multiply *fundamental* entities beyond necessity.

[19] Korman (2015) points out that metaphysicians who think that composite entities (e.g., tables) do not exist usually argue for this thesis on the basis of considerations of vagueness, or by citing what they think are plausible metaphysical principles about identity. He doesn't think that parsimony is relevant here. It may be thought that overdetermination problems need to be addressed by invoking parsimony. If Z is determined (in some relevant sense) by X and also by Y, is the right response that a choice must be made, on grounds of parsimony? If it is, then this role for parsimony has no place in the three parsimony paradigms I've described.

A different issue arises when one asks which of two sets of postulates is preferable when they have identical consequences with respect to the subject matter that they are supposed to systematize. The dispute that ensued after the discovery of set-theoretic paradoxes is a case of this sort. Zermelo-Fraenkel set theory was in competition with Russell's theory of types. The former formulation won the day, to some degree owing to its being simpler and easier to use. However, I suggest that that difference doesn't license the conclusion that ZF is true and type theory is false. The relevant razor involves silence, not denial.[20]

This point pertains to metaphysicians' concern with fundamentality. Sider (2011, 2013) thinks that questions about fundamentality pertain not just to the category of objects; he thinks it also concerns the properties and relations that are cited in theories. To use a familiar example (simplified from Goodman, 1965), consider the following two definitions:

> An object is grue at time t precisely when the object is green at t and $t < 2050$ or the object is blue at t and $t \geq 2050$.
>
> An object is bleen at time t precisely when the object is blue at t and $t < 2050$ or the object is green at t and $t \geq 2050$.

Now consider the following two statements:

(5) All emeralds are green.

(6) All emeralds are grue until 2050 and thereafter they are bleen.

These two statements are logically equivalent, but Sider thinks there is a difference between them that is important in metaphysics. Sider's question isn't whether we should be nominalists or Platonists about properties. He is asking which properties "carve nature at its joints." Sider thinks that colors do and grulers do not. Using David Lewis's (1983) terminology, we might mark this distinction by saying that colors (green, blue, etc.) are "natural" properties, whereas grulers (grue, bleen, etc.) are not. Sider's idea is that the task of figuring out what is metaphysically fundamental pertains to properties and relations at least as much as it does to objects.

Something like this does go on in science. For example, special relativity says that simultaneity is not a two-place relation, but involves a third *relatum*, a rest frame. The theory also says that it's a mistake to regard the spatial distance between two events and the temporal distance between those two events as two objective and physically independent facts about the events; rather, it's the space-time distance between events that is said to be real. Fair enough, but here we're appealing to an empirical theory to justify these judgments, and the theory in question has lots of observational confirmation. Nothing like this pertains to the metaphysician's claim that (5) and (6) are different.

Do the inferential methods used in science allow one to compare (5) and (6) and similar pairs of logically equivalent statements for their joint-carving prowess? The

[20] I thank Bruno Whittle for drawing my attention to this example.

answer is *no* if the epistemology used in science is probabilistic. It's a theorem of probability theory that Pr(X|Z) = Pr(Y|Z) and Pr(Z|X) = Pr(Z|Y) if X and Y are logically equivalent. Recall that the three parsimony paradigms described earlier all make use of probability theory. The thesis that (5) gets at what is fundamental while (6) does not floats free from the probabilistic epistemologies that scientists rightly use in their own subjects.[21] Sider (2013, p. 239) claims that "ideologically simpler theories are more likely to be true." If (5) is ideologically simpler than (6), this claim *cannot* be right.

Quine (1970/1986, p. 80) wasn't thinking about parsimony or probability when he claimed that there is no substantive question concerning which of several axiomatizations of logic or set theory is best if the alternatives have exactly the same consequences. He grants that some may be more elegant or intuitive, but he doesn't think that is epistemically relevant. Quine's naturalism is doing the work here, and the naturalism he is espousing is an instance of MN_p.

13.4 Empirically Equivalent Theories and Identifiability

Empirical equivalence is a weaker form of equivalence than logical. Two theories are empirically equivalent precisely when it's impossible for observational data to discriminate between them. Here "impossible" means something much weaker than logically impossible. I won't try to nail down what the precise modal concept is, but will just note that empirical equivalence goes far beyond the modest fact that our *present* data fail to discriminate between the two theories.

If scientists rightly use parsimony considerations to decide which of two empirically equivalent scientific theories is true, then MN_p opens the door for philosophers to use parsimony to decide between empirically equivalent philosophical theories. Maybe this idea applies to theories that make no predictions about observations at all; maybe we can say that they are vacuously empirically equivalent, and maybe parsimony can be brought to bear on those discrimination problems as well.

I say yes to these "maybe's," but lots of work is needed to make them into something you can hang your hat on. Consider the three parsimony paradigms I've described. The likelihood paradigm involves cases in which the simpler of two hypotheses has the higher likelihood, but this means that the two hypotheses confer different probabilities on the observations. The model selection paradigm involves fitting models with adjustable parameters to observations, and then thinking about which of the fitted models will do better in predicting new observations. In this context, it is possible for two models to fit present data equally well and for the difference in parsimony to be epistemically relevant to estimating which

[21] It may be thought that deterministic theories in science don't require an epistemology that is probabilistic. Not so! Observational error is always possible, and the standard way to model this is by using probabilities.

fitted model will be more predictively accurate. The question is whether competing philosophical theories contain adjustable parameters that can be estimated from observations. But more importantly, if you *know* that two models are empirically equivalent, you shouldn't be using AIC to figure out which of them will be more predictively accurate. The third parsimony paradigm involves assigning prior and posterior probabilities to hypotheses. The razor of silence is on firm philosophical footing here when one theory is logically stronger than the other. However, if the philosophical theories you wish to evaluate are mutually incompatible, and you wish to say which of them is true or which is more probable, the razor of silence is of no help.

There is an additional reason to be skeptical of the epistemic relevance of parsimony in scientific contexts when the competing theories considered are empirically equivalent. It involves the statistical idea of *identifiability*. To get the rough idea of what this means, consider a linear model for the variables x and y:

(LIN) $y = sx + i$

Here s is the slope and i is the y-intercept. This model is identifiable if you have 2 or more data points, meaning that you can use any such data set to derive unique maximum likelihood estimates of the two parameters. And if LIN is true, your estimates for s and i will converge on their true values as your set of data points gets larger and larger (provided that you assemble your data by random sampling from possible x values).

Now let's consider a different model of how x and y are related. To construct this model, I'll introduce two new parameters, u and v, with the stipulation that $s = u + v$. Here's the result:

(LIN*) $y = (u + v)x + i$

This model is *not* identifiable given n data points (for any finite n). True, you can use your data to derive a maximum likelihood estimate of the *sum* of u and v, but unique estimates for *each* of u and v are impossible.

Do scientists say that LIN* should be rejected because it is unparsimonious, and should they say this? My answer is a double negative. I think scientists will fault LIN* for not being applicable to real-world systems. It is *pointless* to "split" the parameter s (whose value can be estimated) into the sum of two parameters (neither of which can be estimated).[22] However, that is no reason to conclude that LIN* is false. Indeed, if $s = u + v$, no reasonable principle of inference should lead you to conclude that LIN is true and LIN* is false. The fact that LIN has fewer adjustable parameters than LIN* cuts no ice.

[22] This point about identifiability is worlds away from verificationism, which claims that untestable propositions are meaningless.

A similar line of reasoning applies to a second and less trivial example. Suppose you want to model the relative velocity of two objects that are moving at constant velocity with respect to each other during a given time interval. You could write the model this way:

(RV) RelVel(a,b) = c RelVel(b,a) = −c

RV has one adjustable parameter. A competing model might be formulated that talks about the absolute velocity of each object:

(ABS) AbsVel(a,$) = c_1 AbsVel(b,$) = c_2

Here $ is absolute space. The ABS model has two adjustable parameters, and they cannot be estimated from data. Were the values of c_1 and c_2 known, the value of c could be computed. In contrast, were the value of c known, infinitely many pairs of values for c_1 and c_2 would be ruled out, but infinitely many would remain.

Should you conclude that ABS is false and that RV is true because RV is more parsimonious? This conclusion can't be justified by appeal to AIC. ABS is not identifiable; its parameters can't be estimated from data, so the AIC score of ABS is not defined. RV is not in that uncomfortable position. Scientists should fault ABS for failing to be applicable to real-world systems. However, that does not mean that the model is *false*. This point about identifiability involves no commitment to AIC or to the instrumentalist philosophy with which AIC is sometimes associated.

The story changes if you have empirical evidence that there is no such thing as absolute space. Then it's not just that you can't estimate the parameters in ABS; you additionally have evidence that a *relatum* mentioned by the model does not exist. In this case, no wielding of the razor of denial is needed. On the other hand, if absolute space can explain observations that a purely relational view of space cannot (as Newton thought in connection with his bucket), that is a point in favor of the hypothesis; again, the fact that ABS is less parsimonious doesn't matter.

13.5 Unification in Science and Metaphysics

Scientists often praise scientific theories for their unifying power, and metaphysicians often do the same with respect to their theories. Newton's laws of motion unified terrestrial and celestial motion. David Lewis (1983) argues that his idea of "natural properties" provides a unifying treatment of the distinction between intrinsic and extrinsic properties, the Kripkenstein puzzle about meaning, and laws of nature. This was not a one-off appeal to unification on Lewis's part. He also argued that realism about possible worlds provides a unifying account of properties, propositions, conditionals, causation, and modality (Lewis, 1986). Are invocations of unification in metaphysics on the same footing as appeals to unification in science? In the latter, unification is supposed to be epistemically relevant. Is it also relevant in metaphysics, and does the argument for its epistemic relevance in science carry over to metaphysics?

I did not mention unification in Sect. 13.1, but it is there nonetheless. The common ancestry hypothesis I considered there is more unifying that the separate ancestry hypothesis is.

According to the former, there is a token evolutionary process (the one that leads to the most recent common ancestor of the two species) that helps explain the fact that humans and chimpanzees both have tail bones. No such unifying process is invoked by the hypothesis of separate ancestry. As for the NULL and DIFF models about the two fields of corn, they can be rewritten so that they exemplify the distinction between unification and disunification:

(UNI) $h_1 = h_2 = v$
(DIS) $h_1 = v_1$ and $h_2 = v_2$

The UNI hypothesis is unifying because its one adjustable parameter applies to both fields of corn. DIS is disunifying because it assigns different parameters to the different fields. Unification in the comparison of UNI and DIS has the same rationale that parsimony possesses in the comparison of NULL and DIFF.

In the second parsimony paradigm, the unifying power of the common ancestry hypothesis is epistemically relevant by virtue of its connection with the law of likelihood. In the third parsimony paradigm, the unifying power of UNI is epistemically relevant because of its connection with AIC. However, neither of these rationales for the epistemic relevance of unification applies to Lewis's theories.

Suppose, for example, that his realism about possible worlds entails true propositions pertaining to each of three subjects; call those propositions X, Y, and Z. To apply the Law of Likelihood here, you need a competing hypothesis. Suppose it's the disunifying philosophical theory D, which is a three-fold conjunction; the first conjunct entails X, the second entails Y, and the third entails Z. If both Lewis's theory and the disunifying competitor D entail the propositions in question, their likelihoods are the same (=1). Here the warning issued at the end of Sect. 13.1.2 comes into play.

If AIC is to apply to Lewis's theories, they must have adjustable parameters whose values be estimated. One worry about both theories is that they merely *accommodate* our philosophical intuitions. The concern is that "naturalness" of properties and "similarity" among possible worlds are not sufficiently characterized. True, there are *examples* of each that point to the relevant distinction (e.g., green is natural but grue is not), but plausible examples are no substitute for a general criterion that can be checked against intuitions.

One more detail about AIC pertains to Lewis's modal realism. As mentioned, predictive accuracy is not the same as truth or probable truth. This means that even if Lewis's theory of possible worlds managed to have a high AIC score (compared to the score of some competing theory), that would not justify the claim that it is true. The usefulness of the theory would be vouchsafed, but that finding would leave open whether one should say that the theory is true or that it is merely a useful fiction (aka "a model").

13.6 Quinean Epistemological Holism

Using parsimony as a tool for evaluating philosophical theories does not require MN_p, but MN_p provides a comfortable home for that use of Ockham's razor. Similarly, MN_p does not require Quine's epistemological holism, but Quinean holism provides a comfortable home for MN_p. Transitivity applies here; Quine's epistemological holism provides a comfortable home for using parsimony as a philosophical tool. Indeed, this isn't a mere possibility; philosophers have embraced Quinean holism, and have taken that framework to provide a green light for using Ockham's razor in their subject. Ted Sider's (2011, 2013) work exemplifies this approach.

There are different sorts of epistemological holism, depending on what epistemic concept one is considering. Let's begin with confirmation. Confirmational holism comes in two forms. The first is *distributive* holism; it says that when an observation confirms a whole theory (which can be regarded as a conjunction), it also confirms the conjuncts. The second is *nondistributive* holism; it says that observations confirm whole theories only; they never confirm their constituent conjuncts (Sober, 1993, 2004). Quine was a confirmational holist, but which sort of holist was he?

Quine used his holism to oppose Carnap's idea that theories typically include propositions that differ in their epistemic status. Carnap, like other logical positivists and logical empiricists, thought that scientific theories often include conventions. These conventions are justified by their usefulness, not by there being empirical evidence on their behalf. For Carnap (1950a), the claim that physical objects exist and the claim that electrons exist are different in kind. The former answers an external question; the latter answers a question that is internal.[23] Belief in physical objects is useful, but that's not evidence that such things exist. However, once you assume that physical objects exist, you can muster empirical evidence for the existence of electrons. For Quine (1953, p. 41), this distinction involves an untenable dualism: "our statements about the external world face the tribunal of sense experience not individually but only as a corporate body." Some of those beliefs may be more "central" than others (Quine & Ullian, 1970a, b), but this is a difference in degree, not a difference in kind. Quine casts the analytic/synthetic distinction and the distinction between *a priori* and *a posteriori* into the outer darkness. He thinks we have the strongest possible reason[24] to believe that physical objects and numbers exist. The reason is that these postulates figure in our highly confirmed physical theories; the postulates inherit their justification from the highly confirmed theories in which they occur. Indeed, these postulates are *indispensable*

[23] Carnap (1950a) tied his epistemological claim about the difference between answers to internal questions and answers to external questions to linguistic issues. I think his epistemological idea should be separated from his linguistic formulation.

[24] This reason isn't like the reason that Pascal offers for believing in God. It is epistemic, not prudential, though perhaps Quine's pragmatism may require him to regard this as yet another untenable dualism.

to those theories; the theories couldn't even be formulated without assuming that physical objects and numbers exist.[25] Quine's holism is distributive.

Epistemological holism, in both its distributive and its nondistributive form, is mistaken when it is formulated as a claim about confirmation. The flaw in the distributive version can be seen by considering a principle that Hempel (1945) formulated:

> **the special consequence condition of confirmation:** If O confirms H, and H entails C, then O confirms C.

Hempel (1945) called this a "condition" since he thought it was a condition of adequacy that an explication of the concept of confirmation must satisfy.

Hempel could not have known when he first published this idea that Carnap (1950b, p. 397; for discussion see Salmon 1975) would prove that this condition is false, if "O confirms H" means that O raises H's probability.[26] This definition of "incremental confirmation" is the standard one that has been adopted by Bayesian philosophers. Here's a simple example that shows that the special consequence condition is wrong. You are dealt a card at random from a standard deck. Here are the observation (O), the hypothesis (H), and the consequence (C) to consider:

O = the card before you is red.
H = the card is the Jack of hearts
C = the card is a Jack

The relevant probabilities are these:

$$\Pr(O) = 1/2 \quad \Pr(H) = 1/52 \quad \Pr(C) = 1/13$$
$$\Pr(H|O) = 1/26 \quad \Pr(C|O) = 1/13$$

[25] It may seem too crude to formulate confirmational holism as a claim about the conjuncts in a conjunction. Can the pure mathematics used in an empirical theory be separated from the empirical claims that the theory make? Surely the existence of pure mathematics is enough to guarantee that you can isolate the pure mathematics. But how to formulate the rest of the theory? Maybe a reasonable beginning can be made by having "numbers exist and have properties X" be one conjunct and "if numbers exist and have properties X, then ..." as the other. In the end, a conjunctive formulation may not be fully satisfactory, but epistemological holism does not require it. Perhaps it's better to drop talk of conjunctions and their conjuncts, and simply say that confirmational holism maintains that whatever confirms a theory thereby confirms its consequences.

[26] Hempel mentions Carnap's result in the 1964 Postscript he wrote for the paper, which appeared in his widely read book (Hempel, 1965).

O raises H's probability, but O does not raise C's.[27,28]

The postulate that physical objects exist and the postulate that numbers exist are used in theories we think are true, but they also are used in theories we think are false. The thought that both postulates are confirmed because they are indispensable in theories we think are true ignores the fact that they are equally indispensable in theories we think are false. Ronald Reagan was once called "the Teflon President" because criticisms never stuck to him, though praise did. Distributive holists seem to think that "numbers exist" and "physical objects exist" have a similar Teflon coating (Sober, 2000). Confirmational holists must address the following question: if confirmation distributes, why doesn't disconfirmation?

Since the argument just given against epistemological holism uses the Bayesian concept of confirmation, the question arises of whether holism is a good principle when applied to other epistemological concepts. For example, consider the Law of Likelihood (Sect. 13.1.2) as it applies to two skeptical puzzles. The first is the evil demon problem. You now seem to see a printed page before you; call this proposition E. Does E favor the hypothesis that there is a printed page in front of you over the hypothesis that your experience is being caused, not by a printed page, but by an evil demon? The law of likelihood says that the answer is *no*:

[27] Moore's (1939) proof of the external world connects with this point about the special consequence condition. Consider the following argument (which isn't exactly Moore's):

> The experiences I now am having provide strong confirmation for the proposition that I have a hand.
>
> If I have a hand, then physical objects exist.
>
> The experiences I now am having provide strong confirmation for the proposition that physical objects exist.

The argument is invalid.

[28] Bayesianism rejects holism as it applies to incremental confirmation, but there is another part of Bayesianism that may seem to be on Quine's side in his disagreement with Carnap. If a theory T entails the postulate (P) that physical objects exist, then $Pr(P \mid X) \geq Pr(T \mid X)$, for any proposition X, provided that the two probabilities are well-defined. This means that if your evidence tells you that T has a high probability, that same evidence tells you that P's probability is at least as high. This point about proposition P holds for every other proposition that T entails. This sounds like epistemological holism on steroids, but consider this: It doesn't matter whether X is empirical evidence for theory T; X could be a tautology, or evidence against T, or an empirical statement that is evidentially irrelevant to T. These points indicate that the inequality does nothing to show that *evidence for* a theory thereby provides *evidence for* the theory's consequences. Carnapians can grant the inequality and still insist that conventions are different in kind from empirical statements. A convention (C) in the Carnapian sense resembles a tautology in that $Pr(C \mid O) = Pr(C \mid notO)$, where O is an observation statement (again assuming that both probabilities are well-defined).

Carnap (1950b) famously distinguished two senses of confirmation — incremental confirmation and absolute confirmation. The latter might better be called "affirmation." E provides the strongest possible affirmation of H when E entails H; the degree to which E affirms H declines as Pr(H|E) declines.

Pr(E | there is a printed page before you & your visual system is now reliable) = Pr(E | there is no printed page before you & an evil demon is causing you to have the experience you'd have if there were a printed page before you).

Reichenbach (1958) makes a similar point about physical geometry and the forces that act on physical objects. His thesis can be represented as a likelihood equality:

Pr(O | space is Euclidean & a universal force is at work) = Pr(O | space is nonEuclidean & there are no universal forces)

Here O describes a set of measurements – e.g., measurements of the angles formed by connecting three points on different mountain tops by light rays.[29]

Setting aside the question of whether these two likelihood equalities are true, I suggest that these equalities do not entail the claims that distributive holism licenses when applied to the law of likelihood, namely these:

Pr(E | there is a printed page before you) = Pr(E | there is no printed page before you)
Pr(E | your visual system is now reliable) = Pr(E | an evil demon is causing you to have the experience you'd have if your visual system were reliable).
Pr(O | space is Euclidean) = Pr(O | space is nonEuclidean)
Pr(O | a universal force is at work) = Pr(O | no universal force is at work)

Distributive holism also fails in the context of model selection criteria like AIC.[30]

At the end of "Two Dogmas of Empiricism," Quine says that he is "impressed also, apart from prefabricated examples of black and white balls in an urn, with how baffling the problem has always been of arriving at any explicit theory of the empirical confirmation of a synthetic statement" (Quine, 1953). Quine was aware of Carnap's *Logical Foundations of Probability*; indeed, he refereed the book for University of Chicago Press (Quine, 1946, pp. 399–402). Unfortunately, Quine inadvertently erected his holistic epistemology on the faulty foundations of the special consequence condition.

[29] Sider (2011, pp. 38–43) discusses this example.

[30] What about nondistributive holism? Skeptical problems concerning evil demons and universal forces are poster children for that form of holism. However, scientific discrimination problems are very often not like this. It is an important goal of science, frequently attained, to break a conjunctive theory into its conjuncts and assemble evidence that bears on some of those conjuncts without saying anything about the others. Here's an example. Consider the hypothesis (H) that you have the covid virus. This hypothesis, by itself, does not tell you what the probability is that your covid test result will come out positive, but it does so when an auxiliary assumption (A) is added, namely that that the test procedure has a given level of reliability. This is a probabilistic version of Duhem's (1914) thesis. The point of relevance to nondistributive holism is that the conjunction H&A can be pulled apart, with A tested without reference to H (Sober, 2004). For another example, consider Darwin's theory of evolution, which includes claims about natural selection and common ancestry. Darwin took adaptive traits to be strong evidence for the former, but not for the latter (Sober & Steel, 2017).

13.7 Conclusion

I think metaphysicians are often whistling in the dark when they claim that their use of parsimony has the same epistemic standing as the use of parsimony in science. Sometimes they seem to think they are safe in wielding the razor because the epistemic role of parsimony in science is an ineffable mystery. These rumors of ineffability are greatly exaggerated.

I do not claim that the three parsimony paradigms I've describe here (and in Sober, 2015 in more detail) are exhaustive. Schulte (1999) and Kelly (2007) argue that there is a pragmatic justification for testing theories in order of their simplicity, and there are other ideas afloat in probability and statistics that are thought to demonstrate parsimony's epistemic relevance. An example is the "VC criterion" named for Vladimir Vapnik and Alexey Chervonenkis (on which see Vapnik, 2000), which is influential in the theory of machine learning. What I do claim is that the three parsimony paradigms I've described, along with the distinction between the razor of silence and the razor of denial, provide a starting point for tackling the problem of whether the principle of parsimony used in philosophical reasoning has the same justification that the principle has in the context of scientific reasoning. This is a worthwhile question in meta-philosophy.

Metaphysicians using Quine's (1951) terminology often distinguish between ontological and ideological parsimony. The former is relevant to evaluating competing claims about what exists; the second is where metaphysical questions about fundamentality make their appearance. I have expressed my skepticism about the claim that parsimony, as it is used in science, is relevant to discriminating between observationally equivalent philosophical theories; it doesn't matter whether one of those theory has greater ideological or ontological parsimony than the other. Here the distinction between the razor of silence and the razor of denial is key.

That distinction is especially pertinent when metaphysical problems are formulated like this: "We all agree that Fs exist. The question is whether G's do too." This formulation gives the impression that the hypotheses to consider are "Fs exist" and "F's and G's both exist." As noted, the former can't be less probable than the latter, but that leaves open whether "Fs exist and G's do not" is more probable than "F's and G's both exist." It is the latter pair of claims, not the former, that usually are the focus of philosophical attention. Nominalists about properties and nihilists about composite objects are nay-sayers, not skeptics.

I think there is no global and unconditional defense of the epistemic relevance of parsimony in science, but that doesn't mean that parsimony is epistemically irrelevant. Here I'm extracting a lesson from the debate between I. J. Good (1967, 1968) and Carl Hempel (1967) concerning the ravens paradox. Good was right; there is no general and unconditional justification of the Nicod criterion – that a generalization of the form "All As are B" is confirmed by observing an object that is both A and B. However, that doesn't mean that observing a black raven fails to confirm the hypothesis that all ravens are black. Confirmation is a three-place relation between an observation, a hypothesis, and background assumptions.

The bearing of parsimony on hypotheses likewise involves a third relatum. Hand-wringing about parsimony's lack of justification in science comes from looking in the wrong places.

Nothing said here precludes the possibility that parsimony is epistemically relevant in metaphysics. My complaint concerns the assumption that parsimony is epistemically relevant in metaphysics because parsimony is epistemically relevant in science. Maybe metaphysics has its own epistemic ground rules. These need to be spelled out if the crutch of science is not available.

Acknowledgments My thanks to Dylan Beshoner, Jeremy Butterfield, Richard Creath, Jordan Ellenberg, Matthew J. Maxwell, William Roche, Casey Rufener, Alan Sidelle, Ted Sider, Mike Steel, Bruno Whittle, and Shimin Zhao for useful discussion.

References

Akaike, H. (1973) "Information theory and an extension of the maximum likelihood principle." In: B. Petrov and F. Csaki (eds.) Second International Symposium on Information Theory. Akademiai Kiado, pp. 267–281

Bennett, K. (2017). *Making things up*. Oxford University Press.

Brenner, A. (2017). Simplicity as a criterion of theory choice in metaphysics. *Philosophical Studies, 174*, 2687–2707.

Carnap, R. (1950a). Empiricism, semantics, and ontology. *Revue Internationale de Philosophie, 4*, 20–40. Reprinted in *Meaning and necessity — A study in semantics and modal logic*, 2nd edition, 1956, Chicago: University of Chicago Press.

Carnap, R. (1950b). *Logical foundations of probability* (2nd ed.). University of Chicago Press. 1963.

Cowling, S. (2013). Ideological Parsimony. *Synthese, 190*, 3889–3908.

Draper, P. (1989). Pain and pleasure — An evidential problem for theists. *Noûs, 23*, 331–350. Reprinted in D. Howard-Snyder (ed.), *The Evidential Argument from Evil*. Bloomington, IN: Indiana University Press, 1996, pp. 12–29.

Duhem, P. (1914). *The aim and structure of physical theory* (P. Wiener, Trans.). Princeton: Princeton University Press, 1954.

Field, H. (1980). *Science without numbers*. Princeton University Press.

Forster, M., & Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1–36.

French, S. (2014). *The structure of the world — Metaphysics and representation*. Oxford University Press.

Good, I. J. (1967). The white shoe is a red herring. *British Journal for the Philosophy of Science, 17*(4), 322.

Good, I. J. (1968). The white shoe qua red herring is pink. *British Journal for the Philosophy of Science, 19*(2), 156–157.

Goodman, N. (1965). *Fact, fiction, and forecast*. Bobbs-Merrill.

Hacking, I. (1965). The Logic of Statistical Inference. Cambridge University Press.

Harman, G. (1977). *The nature of morality*. Oxford University Press.

Hempel, C. (1945). Studies in the logic of confirmation. *Mind, 54*(1–26), 97–121.

Hempel, C. (1965). Studies in the logic of confirmation. In *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 3–51). Free Press.

Hempel, C. (1967). The white shoe—No red herring. *British Journal for the Philosophy of Science, 18*(3), 239–240.

Huemer, M. (2009). When is parsimony a virtue. *The Philosophical Quarterly, 59*, 216–236.
Kelly, K. (2007). A new solution to the puzzle of simplicity. *Philosophy of Science, 74*(5), 561–573.
Korman, D. (2015). Fundamental quantification and the language of the ontology room. *Noûs, 49*, 298–321.
Kriegel, U. (2013). The epistemological challenge of revisionary metaphysics. *Philosophers' Imprint, 13*(12), 1–30.
Lakatos, I. (1978). In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programmes. Philosophical Papers* (Vol. 1). Cambridge University Press.
Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy, 61*, 343–377.
Lewis, D. (1986). *On the plurality of worlds*. Blackwell.
Malament, D. (1982). Review of Harty Field's Science without numbers. *Journal of Philosophy, 79*(9), 523–534.
Moore, G. E. (1939). Proof of an external world. *Proceedings of the British Academy, 25*, 273–300.
Quine, W. (1946). Quine to University of Chicago Press 1946-10-29. In R. Creath (Ed.), *Dear Carnap, Dear Van: The Quine-Carnap correspondence and related work by W.V. Quine and Rudolf Carnap*. University of California Press. 1990.
Quine, W. (1951). Ontology and ideology. *Philosophical Studies, 2*, 11–15.
Quine, W. (1953). Two dogmas of empiricism. In *From a logical point of view*. Harvard University Press.
Quine, W. (1970a). *The web of belief*. Random House.
Quine, W. (1970b). *Philosophy of logic*. Harvard University Press.
Rawls, J. (1971). *A theory of justice*. Harvard University Press.
Reichenbach, H. (1956). *The direction of time*. University of California Press.
Reichenbach, H. (1958). *The philosophy of space and time*. University of California Press.
Salmon, W. (1975). Confirmation and relevance. In G. Maxwell & R. Anderson (Eds.), *Induction, probability and confirmation. Minnesota studies in philosophy of science* (Vol. 6, pp. 3–36). University of Minnesota Press.
Schaffer, J. (2015). What not to multiply without necessity. *Australasian Journal of Philosophy, 93*, 644–664.
Schulte, O. (1999). Means-ends epistemology. *British Journal for the Philosophy of Science, 50*, 1–31.
Shapiro, L., & Sober, E. (2007). Epiphenomenalism – The do's and the don'ts. In P. Machamer & G. Wolters (Eds.), *Thinking about causes* (pp. 235–264). University of Pittsburgh Press.
Sider, T. (2011). *Writing the book of the world*. Oxford University Press.
Sider, T. (2013). Against parthood. In K. Bennett & D. Zimmerman (Eds.), *Oxford studies in metaphysics* (Vol. 8). Oxford University Press.
Sober, E. (1988). *Reconstructing the past – Parsimony, evolution, and inference*. Cambridge University Press.
Sober, E. (1993). Mathematics and indispensability. *Philosophical Review, 102*, 35–58.
Sober, E. (2000). Quine's two dogmas. *Proceedings of the Aristotelian Society, 74*, 237–280.
Sober, E. (2004). Likelihood, model selection, and the Duhem-Quine problem. *Journal of Philosophy, 101*, 1–22.
Sober, E. (2009a). Absence of evidence and evidence of absence – Evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies, 143*, 63–90.
Sober, E. (2009b). Parsimony arguments in science and philosophy – A test case for naturalism p. *Proceedings and Addresses of the American Philosophical Association, 83*(2), 117–155.
Sober, E. (2015). *Ockham's Razors – A user's manual*. Cambridge University Press.
Sober, E. (2018). *The design argument*. Cambridge University Press.
Sober, E., & Steel, M. (2017). Similarities as evidence for common ancestry – A likelihood epistemology. *British Journal for the Philosophy of Science, 68*, 617–638.
Sotos, J. G. (1991/2006). *Zebra cards – An aid to obscure diagnoses*. Mount Vernon Book Systems.

Thomasson, A. (2015). *Ontology made easy*. Oxford University Press.
Vapnik, V. (2000). *The nature of statistical learning theory*. Springer.
Vassend, O., Sober, E., & Fitelson, B. (2017). The philosophical significance of Stein's paradox. *European Journal for Philosophy of Science, 7*, 411–433.
Willard, M. (2014). Against simplicity. *Philosophical Studies, 167*, 165–181.
Wright, E., Levine, A., & Sober, E. (1992). Marxism and methodological individualism. In *Reconstructing Marxism – Essays on explanation and the theory of history*. Verso Press.
Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *Philosophical Magazine, 42*, 369–390.

Chapter 14
Levels, Kinds and Multiple Realizability: The Importance of What Does Not Matter

James Woodward

Abstract This essay discusses the notions of levels, multiple realizability and kindhood from the perspective of an interventionist account of causation and explanation. The notion of level on which I will focus is explanation-based—the idea is that sometimes the fine-grained details of a system do not matter or do not matter much for the explanation of certain coarse-grained features of its behavior. When this is the case, we often regard those fine-grained details as at a "lower-level" than the "upper-level" coarse-grained behavior. Philosophers have developed many different versions of multiple realizability (MR) but I argue that the most defensible account is in terms of the explanatory irrelevance of lower-level details characterizing realizers to what is realized at an upper level. This leads to a somewhat novel understanding of MR that drops certain implausible claims about that notion. Another, related theme is that levels and MR should be understood in terms of relations between variables rather than kinds. I argue that the common philosophical claim that laws and causal generalizations relate kinds (rather than variables) is a source of many confusions.

14.1 Introduction

Notions of level (of organization, explanation, composition or realization, even of "being") are widespread both in science and philosophy, as is talk of "upper" and "lower" levels and (in some quarters), contrasts between "the fundamental" as opposed to "non-fundamental" levels. At the same time, some (both in philosophy

Many thanks to Larry Shapiro for extensive comments on an earlier draft.

J. Woodward (✉)
University of Pittsburgh, Pittsburgh, PA, USA
e-mail: jfw@pitt.edu

S. Ioannidis et al. (eds.), *Levels of Reality in Science and Philosophy*,
Jerusalem Studies in Philosophy and History of Science,
https://doi.org/10.1007/978-3-030-99425-9_14

and outside of it) regard such notions as misleading and a potent source of confusion (Eronen, 2015).

My focus in this essay will be on levels of explanation and their relationship to multiple realizability and to notions of kindhood. The explanation-related notion of levels on which I will focus involves (very roughly) the idea that sometimes the lower-level details of a system do not matter or do not matter much to some aspects of its upper-level behavior. When this is the case those aspects can be satisfactorily explained without reference to lower-level details. It is this idea that I refer to in my title—what doesn't matter is important in understanding levels of explanation, multiple realizability and related ideas.[1] As we shall see, this idea that requires considerable unpacking—I attempt to accomplish this in terms of a notion that I call conditional causal independence.

This conditional independence notion of level is just one of a number of different notions of level discussed in the philosophical and scientific literature; my reasons for focusing on it will become clearer below. Although I agree with critics that level talk can sometimes mislead, I think the general idea that there are different possible levels of explanation for various kinds of behavior is, in the sense I will describe, very defensible. Indeed it is central to understanding how scientific investigation can get a fruitful grip on the world. On this picture which levels are most appropriate for understanding the behavior of a system is not a matter to be decided *a priori* but instead will depend on the empirical details of that behavior—in particular, on specific empirical facts about which factors matter and do not matter to the behavior in question. I thus reject the idea that there is, independently of such specific empirical considerations, a single most fundamental level which is most appropriate for explaining all varieties of system behavior or which provides explanations that are deeper or better than explanations at other levels.

Discussions of levels of explanation are closely linked, both in philosophy and elsewhere to the notion of *multiple realizability* (hereafter MR)—which I will understand as the idea that the same upper-level behavior (or conformity to the same upper-level generalizations) can be "realized" by systems that differ in their lower-level details. I thus view MR as a way of capturing the idea that sometimes lower-level details do not matter to upper-level behavior. When understood in this way, claims about multiple realization are very closely related to the notion of level as conditional independence referred to in the previous paragraph.

Philosophical arguments frequently appeal to multiple realizability in support of the claim that upper-level generalizations can figure in legitimate explanations. At the same time, the notion has been attacked on various grounds, both conceptual and empirical. (See e.g., Kim, 1993; Polger & Shapiro, 2016). My view is that, properly understood MR is an important and coherent notion and that as an empirical matter there are a number of systems that exhibit this feature. At the same time

[1] The importance of irrelevance or what doesn't matter is also emphasized in recent work by Robert Batterman—see, e.g. Batterman, 2021. I've been much influenced by Batterman's ideas but don't mean to suggest that we agree in all respects.

I think that many of the most influential discussions of MR have encrusted it with various features that have been major sources of confusion. These include arguments that understand multiple realization in terms of different lower-level *kinds* (or properties—I will use these somewhat interchangeably in what follows) realizing the same upper-level kind (or property) and, along with this, claims that certain generalizations are illegitimate because the kinds or properties that figure in them are defective in some way—for example, because they involve "disjunctive" properties. My contrary view is that thinking of relations between levels in terms of relations between kinds is in most cases not very fruitful and that the whole debate around the legitimacy or not of disjunctive properties is confused, roughly because it rests on a misunderstanding of how causal generalizations work. Part of my goal in what follows thus will be to propose a framework that allows us to understand levels of explanation and multiple realizability without taking on board some mistaken commitments that have often been associated with these notions. This alternative framework understands relations between levels and multiple realization as having to do with relations among *values of variables*, rather than with relations between kinds or properties. As we shall see, this also leads to a different way of understanding what the autonomy of upper-level science generalizations consists in than is standard in philosophical discussion. I emphasize that this framework departs in important ways from ways of thinking about MR and related notions that are standard in the philosophical literature—my intent is not to carry out a discussion within this standard framework (or to fully "capture" standard ways of thinking about MR) but to replace this framework with one that I think is better.

The rest of this essay is organized as follows. Section 14.2 contrasts two different notions of level, and defends my subsequent focus on one of these notions, which has to do with conditional independence, rather than a notion organized around compositional considerations. Section 14.3 characterizes the notion of conditional independence more precisely and connects it with notion of a *variable*. Variables are structures like quantities or magnitudes—mass, charge etc.—that can assume different values, with the limiting case being that of a binary variable, which can take just one of two possible values. To avoid continual pedantic qualification I will usually use the word "variable" to describe what is in the world that corresponds to terms like "mass" etc. but in some cases it will be clear that I am talking about the terms themselves—the context will sort out which is intended.[2] I argue that variables, as opposed to philosophical ideas about kinds and properties, are the right notions for understanding conditional independence and multiple realizability. Section 14.4 discusses the notion of a causal generalization and the role of variables in these. Section 14.5 connects the notion of conditional independence to multiple realizability. Sections 14.6 and 14.7 argue that because multiple realizability should not be understood in terms of relations among kinds, concerns about disjunctive kinds or properties in discussions of multiple realizability are misplaced. Section

[2] A bit more housekeeping: I will use italic capital letters (*X*, *Y*) to describe variables and capital letters without italics (A, B) as names of things or kinds of things.

14.8 defends the notion of multiple realizability in response to criticisms by Kim and by Polger and Shapiro. Section 14.9 concludes with a discussion of the implications of the conditional independence notion of level for what might reasonably be meant by the "autonomy" of "upper-level generalizations". Here I also argue that some standard philosophical arguments attempting to show that upper-level explanations are "better" than lower-level explanations of the same explananda are both unsuccessful and unnecessary.

14.2 Two Notions of Level and Their Relations

Levels as Compositional Even if we focus just on levels of explanation (as opposed to other notions of level) different thinkers have regarded different notions as most central. Some have focused primarily on compositional or mereological notions of level. Compositional notions of level apply most straightforwardly to *things* or at least to entities that are thing-like (as opposed to quantities or magnitudes that are described by variables). In particular, as I will understand composition-based notions, they attempt to understand levels in terms of part/whole relations, which are most naturally thought of in terms of relations among thing-like entities.[3] Thus we have the familiar idea that protons, neutrons and electrons are at a different and *lower* level than atoms (because they are "parts" of or constituents of atoms), atoms are at a lower level than molecules, molecules are at a lower level than cells, cells are at a lower level than multicellular organisms and so on. As these examples suggest, this notion of level can be thought of as generating a sort of hierarchy or partial order with wholes at one level, having parts that are at a lower level, these in turn having parts at a still lower level and so on.[4]

Insofar as there is a connection between this notion of level and considerations having to do with explanation, it is presumably that level information, understood in terms of part/whole relations, can guide the choice of an appropriate level of explanation. Here there are several natural possibilities. One is that the appropriate level of explanation for wholes (or for the behavior of wholes) is in terms of their parts (perhaps in terms of their most "immediate" parts, if there is some way of making sense of that notion); the behavior of molecules is to be explained in terms of their constituent atoms and so on. Another apparently natural thought is that things

[3] Although things (or at least most of them) have parts, in most cases the notion of parthood does not apply very naturally to properties and still less to variables (understood either as terms or what they describe in the world). Variables are not "composed" of parts that are other variables. This is one of many differences between a compositional notion of level and the notion based on conditional dependence which is framed in terms of variables.

[4] As observed by Humphreys, 1997, this picture can be very misleading insofar as it suggests that the parts of wholes retain the same features that they have when they are not part of the whole. A hydrogen atom in a water molecule does not behave in the same way as an isolated hydrogen atom.

primarily or exclusively interact causally with things at the same level—protons and neutrons interact with each other, atoms interact with other atoms and so on. This leads to the idea that there is something problematic about inter-level causation. (See e.g. Craver & Bechtel, 2007).

Levels as Tied to Conditional Independence Relations The composition-based notion just described contrasts with the conditional independence notion that I prefer. Again, this second notion is naturally expressed in terms of variables and relations among these, rather than relations between things. Informally, the idea is that certain upper-level relationships among variables are insensitive to some range of variation in their lower-level realizers: where *X*, *Y* and *Z* are variables, upper-level *X* may depend (unconditionally) on lower-level *Z,* but *X* does not depend on Z (or depends on Z only in rare cases), given upper-level variable *Y*, where (at least in the simplest cases) *Y* corresponds to some kind of coarse-grained representation of Z. This is what I have in mind when I say that *X* is causally independent of *Z*, conditional on *Y*. Here Z has to do with the lower-level details that "do not matter" which I interpret as meaning that these details do not matter, *given* the information in *Y*. In interesting cases, *Y* will have a much lower dimension or fewer degrees of freedom than Z so that in employing *Y* rather than *Z* we achieve a dimensional reduction with the information in Z that is causally or nomologically relevant to *X* being absorbed into *Y*. (There will be additional information in Z that is irrelevant to X, this may be relevant to other variables besides *X* but it will not be incorporated into *Y*). Thus ideas about dimensionality reduction and degrees of freedom play important roles in this notion of level.

As an illustration, it is approximately true that given the temperature *T* of an ideal gas, further information about the exact details *K* of the kinetic energies of each of the molecules composing the gas is conditionally causally independent of (conditionally causally irrelevant to) the value of other thermodynamic variables like pressure and volume. We think of temperature, pressure and volume as at a different level than *K* and, on the level notion under discussion, this corresponds to the fact that we don't need all of the information in *K* to explain the thermodynamic behavior of the gas—we just need the information in the temperature variable. Similarly in Putnam's well-known example of the square peg that will not fit into the round hole (1975), it would not be correct to say (as some philosophers have) that the lower-level details of the molecular constitution of the peg and the generalizations governing these do not matter at all, but it is arguably correct to say that given or holding fixed the dimensions of the peg and (importantly) its rigidity or inflexibility, these molecular details do not have any further relevance to whether the peg goes into the hole. This is why we can appeal to upper-level facts about the dimensions and rigidity of the peg rather than lower-level molecular details to explain its behavior.

This conditional independence notion of level is also often tied to considerations having to do with scales—spatial, temporal and energetic—and their separation. It is sometimes the case that what happens at one length or energy or time scale is largely conditionally causally independent of what happens at other scales, and

this in turn leads us to think of interactions at one scale as at a different level than interactions at other scales. Conditional independence in such contexts has to do with the fact that the complicated details D at one scale that are relevant to some behavior B can often be summarized in terms of a single or small set of parameters P at another, higher scale which contains all that is relevant to that behavior. In this case D is causally irrelevant to B, conditional on P. For example, some features of a cell (such as the accumulation of protein product) may change over time but very slowly in comparison with some other process (such as transcription factor activities), so that the former may effectively be treated as constant for the purposes of understanding the latter (Alon, 2007). As another illustration, in understanding fluid behavior, it is often appropriate to use the Navier-Stokes equations. These model the behavior of fluids at continuum level length scales—scales at which the fluid can be treated as a continuous medium and at which the fact that real fluids are composed of collections of molecules and are discontinuous at a finer-grained length scale does not matter for various continuum level behaviors—again "does not matter" should be understood as "does not matter conditional on the molecular components of the fluid having certain very generic features, which are summarized in various parameters such as density and viscosity in the Navier-Stokes equations". One thinks of the Navier-Stokes equations and the variables that figure in them as at a different "level" than the variables and generalizations that would be appropriate to characterize the behavior of individual molecules and this difference in levels is, I claim, captured by the conditional independence notion.

The conditional independence and the composition-based notions of level are not completely unrelated, since sometimes size and parthood matter for conditional independence relations. It is true for example that in many respects the components of a cell will interact much more strongly with other components of the same cell than they will with components of other cells, thus justifying modeling strategies for understanding individual cells that ignore the detailed goings on in other cells. Here parthood tracks causal/explanatory relevance at least to some extent.

Nonetheless, the correspondence between the compositional and conditional independence notions of level is very imperfect. Moreover when they point in different directions, the conditional independence notion seems the more important. Consider that both electrons and the nucleus are "parts" of atom. The strong and weak nuclear forces are very important in characterizing the nucleus and in understanding phenomena like fission and various scattering experiments. On the other hand, because these forces are very short range, they are taken to be "effectively irrelevant" to understanding most aspects of the chemical behavior of atoms—this depends instead on the behavior of the electrons surrounding the nucleus and the electromagnetic force. So here parthood per se does not necessarily track what is explanatorily relevant. Again note that "effectively irrelevant" here needs to be understood as meaning something like (approximate) conditional causal irrelevance. If the strong and weak force were sufficiently different, we would not have stable nuclei at all and hence no atoms. So it is not as though the character of these forces does not matter at all. Rather their irrelevance is conditional; as long as the forces are short range, allow for the formation of stable nuclei and meet other

generic conditions, further details about them do not matter. Thus we can think of chemical behavior as at a different level from detailed models of the physics of atomic nuclei and do chemistry largely independently of atomic physics.

14.3 Conditional Independence Characterized: The Role of Variables

Having provided an informal characterization of the conditional independence, I turn now to a more detailed development of the framework that I will use to characterize this notion and from it, multiple realizability. This will also set the stage for some criticisms of alternative ways of understanding MR.

In thinking about how best to characterize MR it is important to keep in mind what it is that we are trying to understand. In my view, our target is a generic fact about our world: that in a range of cases, certain variations in lower-level detail do not matter (or more precisely, matter only conditionally) to various aspects of upper-level behavior. One might imagine a world in which this is not the case—a world in which, to consider a possibility countenanced by Goldenfeld and Kadanoff, 1999, to model the behavior of a bulldozer one has to invoke the details of quantum chromodynamics. They remark that in this case, one would have "model chaos", presumably with the implication that in this case explaining the behavior of the bulldozer would be hopeless since, needless to say, deriving such behavior from QCD is not a serious possibility. It is an important fact that our world is not like this—in explaining upper-level behavior we can often capture the relevant lower-level information in a much smaller and more manageable set of variables and parameters. I see MR and related ideas as of interest because they are part of a framework for understanding this general fact about independence of details. For this reason, I do not think that it is fruitful to follow such writers as Polger and Shapiro, in adopting characterizations according to which it follows virtually automatically that there are few if any instances of MR. We want to understand the "details don't matter" fact, rather than making it difficult to recognize.

I turn next to some remarks about variables and their values since getting clear about these is crucial for understanding conditional causal independence. This will also help to expose confusions in the way in which some philosophers think about causal generalizations. Again, as I will understand the notion, a variable must be capable of taking more than one value. As a limiting case, a variable may be "binary"—that is it may have just two possible values, as when a variable L takes the values 1 or 0 according to whether a light is on or off. But in many cases in science, variables are more "quantitative"—for example, the variable *mass* can take any positive real value. Variables can be used to characterize features of individual objects—a cannon ball weighs 10 kg, has a velocity of 50 m/s at time t and so on. However, a well-behaved variable cannot assign different values to the same object or unit at the same time—the same cannon ball cannot weigh

10 kg and 5 kg. When applied to the same object different values of the same variable "exclude" one another. By contrast when variables are fully distinct—a notion I will say more about below—and no additional constraints are present, there is a sense of "possible" according to which all combinations of values of distinct variables are possible, both for the same individual and different individuals. For example, the position of a particle in a three dimensional space can be characterized by three distinct variables, corresponding to its co-ordinates in each of the three spatial dimensions. The values of these variables can vary independently of each other—specifying the x co-ordinate of the particle does not constrain what its y or z co-ordinates are. Similarly, the position variables for the particle do not constrain the values of the three variables representing the three components of its momentum nor do they constrain the values of these variables for other particles.

The sense of "possible" at work here is roughly the sense captured by the notion of a phase space understood as representation of the different "possible" states of a system. Thus in the case in which we have a system of N particles, the phase space (the points of which correspond to the possible positions and momenta of each particle) will have 6N dimensions, corresponding to 6N distinct variables. Note that "possible" here does not mean causally or nomologically possible. Once we specify laws or causal generalizations characterizing the system this will typically exclude various combinations of values for these variables that are possible in the phase space sense just described, at least over time. That is, the notion of causal or lawful possibility is distinct from and imposed on top of a prior notion of possibility captured by the phase space representation. We thus have the following contrast between the behavior of variables and values of variables: it is not possible for the same variable, describing some object at a time, to take different values but it *is* possible for different variables to take any values in their range, either in describing the same object or different ones. One consequence of this is that we need to be careful to distinguish variables and values of variables, since they behave very differently.[5] This in turn helps to explain why I employ the framework described rather than a more familiar one in which lower-level kinds or properties are described as "realizing" upper-level ones—as we shall see, this last framework elides the distinction between variables and their values.

With this as background, let us now consider how we might represent a relationship between levels where the values of upper-level variables "supervene" on the values of lower-level variables via some non-causal multiple realization or determination relation. (I will say more about the non-causal element here shortly). I will also focus on one of the simplest possibilities in which the relation between the

[5] Although I focus on the difference between variables and their values, this should not be taken to imply that a variable just is a collection or set of its values. Variables typically have much more structure than this. For example, the values of a variable may or may not be measureable on a ratio scale, may or may not permit meaningful notions of addition and so on.

two levels corresponds to a coarse-graining operation such as averaging (of course this is far from the only possibility).[6]

Suppose then that we have two sets of variables $\{U_i\}$ $i = 1, 2, \ldots$ (upper-level variables) and $\{L_j\}$ $j = 1, 2 \ldots$ (lower-level variables). Possible values of each variable are represented by indexed lower cases letters: the possible values of U_i are u_{ik}—u_{11}, u_{12} and so on and the possible values of L_j are l_{jm}. The idea we want to represent is that the values of the U_i supervene or are multiply realized by values of the L_j. We can capture this at least in part by supposing that for each U_i there is a many to one surjective function f that maps a number of different values of the L_j into each value of U_i. (f may map values from different L_j into values of a U_i or, alternatively, different values from the same L_j may be mapped into a value of U_i.—see below). We require that f be a function because we want to exclude the possibility that the same value of L_j is mapped into different values of U_i (This would violate the assumption of supervenience). We require that this function be surjective to capture the standard assumption that every value of each of the U_i is realized by some value (many values) of the L_js. (In other words, there are no values of U_i that have no counterpart in the supervenience base consisting of the L_j). Multiple realizability is captured by the many to one character of the function—that is, we assume that the function is *not* bijective. As noted in Ellis, 2016 and Woodward, Forthcoming, we may think of the values of L_js that are mapped into the same value of U_i as belonging to the same equivalence class; f thus induces a partition of the values of L_j into disjoint equivalence classes each of which corresponds to a single value of U_i.

As mentioned above, it is important that on this characterization multiple realizability is understood as a relation between *values* of variables rather than in terms of "kinds" (or "properties or "things") being "realized" by other kinds (or properties or things). Among other limitations, a framing in terms of kinds does not naturally capture a number of cases of multiple realizability. Suppose that the kinetic energy of each of the component molecules of a gas is represented by a distinct variable K_i, where i ranges from 1 to, say, 10^{23} with these values being mapped into values of the upper-level variable temperature by an averaging function. No particular variable K_i and no particular value of the kinetic energy for an individual molecule is a "realizer" of the variable T. Instead it is the values of the kinetic energies for each of the individual molecules—an individual value for each K_i—that is mapped via a function that averages these values into a single value of the variable T. It is the full set of these K_i values that realizes T. For similar reasons, MR is not to be understood in "compositional" terms—the K_i values that realize T are not "parts" or "constituents" of T.

[6] As noted by Batterman, 2021, there are many examples in which the relation between upper and lower variables is far more complex than simple averaging—for example, the values of the upper-level variables may depend on facts having to do with the connectivity or topology or correlations among values of lower-level variables and specifying these may require information that is "meso-level" and not naturally thought of as part of the lower-level theory.

As illustrated by this example, relations between upper and lower-level variables in real scientific examples often (perhaps typically) involve mathematical operations like averaging, and in many cases, operations on variables at both levels involve not just arithmetic operations but also operations like differentiating, integrating and taking limits. To capture this we need to think in terms of quantitative variables and operations on these. That is, these variables need to have the kind of structure that allows for operations like taking limits, differentiation and so on. The examples in the philosophical literature of qualitative upper-level properties or states or kinds (e.g., pain) being realized by qualitative lower-level properties (e.g., c-fibers firing) do not reflect any of this—indeed, as noted below, in typical discussions it is not even made clear what the possible values of these variables are. Moreover, these properties are not such that one can perform mathematical operations of the sort described above on them.

The relationship between upper and lower that we are attempting to capture is, as I have said, one of supervenience. Like others, I think of this as a non-causal determination relation. But what does "non-causal" mean? I adopt the standard view that causation requires that the variables corresponding to cause and effect be "distinct" from one another. To use David Lewis's example (1986), my saying "hello" cannot cause my saying "hello" loudly, since these variables are not distinct from one another. To explain what "distinctness" involves I appeal to the feature of variables discussed earlier—variables V_1 and V_2 are distinct when all values of V_1 and of V_2 are compossible with each other, where the relevant notion of possibility is the notion of phase space possibility described earlier. In Lewis's example, the variable V_1 corresponding to saying "hello" has (let us suppose) two possible values, $1 =$ saying hello and $0 =$ not saying hello. The variable V_2 might be understood as having three possible values $2 =$ saying hello loudly, $1 =$ saying hello in an ordinary tone of voice and $0 =$ not saying hello. Certain combinations of these variables such as $V_1 = 0$ and $V_2 = 1$ are not possible—in this case for conceptual reasons. This shows that V_1 and V_2 are not distinct. Similarly, certain combinations of values for the lower-level kinetic energy variables and the temperature variable—combinations in which the lower-level variables take values that are mapped into a value for the temperature value $T = t$ but the temperature value takes a distinct value $T = t^* \neq t$ are not possible. (I put aside the issue of the source of this impossibility but I think it is clear that the impossibility is non-causal). This reflects the judgment that the kinetic energies of the component molecules do not cause the temperature of the gas but instead stand in some other determination relation to the temperature.

As remarked above, the notions of MR and supervenience I have characterized are not intended to incorporate various other features that are often associated with these notions in the philosophical literature. For example, Fodor claims that in many cases in which MR is present (as in the relation between psychological and neurobiological kinds) the realizers of the upper-level generalizations may (in fact, likely do) have nothing in common, that there may be no finite list of possible realizers of a sort that we are able to construct and that the upper-level generalizations may be inexplicable from the point of view of the lower-level theory. My characterization of MR does *not* build in any of these assumptions and in fact

I think they are very likely false. Philosophers have also spent a great deal of time worrying about the "metaphysics" of the multiple realization relation. Are individual token events characterized in terms of the upper-level theory "identical" with their realizers on particular occasions or is the relation between these to be understood in some other way—e.g., in terms of constitution or constitutive relevance (whatever that is)? Aside from thinking that "identity"—either of the type or token variety—is the wrong notion for understanding the relation between upper and lower-level theories (on this see below), my characterization of MR makes no claims about these metaphysical issues. My characterization of MR is intended to provide enough formal structure for what follows and nothing more. It turns out that this formal structure is enough for my subsequent discussion of the notions of conditional independence, autonomy and so on to go forward and that is all that I care about.

14.4 Causal Generalizations and Their Realization

As characterized so far, this framework makes multiple realization of values of variables relatively common or "easy". I don't think of this as a bug or problem because I think what really matters is the multiple realization of upper-level causal *generalizations.*[7] I think the extent to which this occurs and how we should think about it is what is really at issue in the debate over multiple realization.

In thinking about this issue we need an account of what it is for a causal generalization (at any level) to be true or correct. I will adopt the account in Woodward (2008). A causal claim of the form X causes Y, with X and Y variables will be true when there are interventions that change values of X that are regularly associated with changes in values of Y. Here "regularly associated" means that for the values of X in question, whenever there are interventions setting those values, the same values of Y or perhaps the same probability distribution for those values follows.[8] When this is the case I will say that there is a stable intervention-supporting relationship linking X to Y. This is an "interventionist" condition for causation.

When X and Y are multiply realized in the sense described above, this immediately gives us a rather strong constraint on what is required for an upper-level causal generalization (or law) linking X to Y to hold. In particular, if there is such a stable generalization linking X to Y ($Y = G(X)$) for some interventions on some values x_i of X, then for every lower-level realization of those x_i, a realization of the associated values (according to G) of Y must hold. In other words for those

[7] Or, more broadly, relationships that are stable or lawful, whether or not we think of them as causal.

[8] Thus for X to cause Y it is not required that interventions on all values of X make a difference for the value of Y but merely that this is true for some values and that when it is true, the interventions on those values have a stable effect on Y. In other words, there is a range of values of X such that for those values, there is a stable response from Y. See Woodward, 2008.

values x_i, the relation between X and Y should satisfy what Woodward 2008 calls "realization-independence" in the sense that this relationship should hold regardless of how X and Y are realized at the lower level. The gas example comes close to fulfilling this requirement. As noted earlier, for a range of values of the upper-level variables, thermodynamic generalizations like $PV = nRT$ will "almost always" hold regardless of how the values of these variables are realized via the lower-level variables characterizing the molecular constituents of the gas. In this connection, note that there are many possible upper-level variables that can be constructed from lower-level variables and are in this sense multiply realized—for example, one might define an upper-level variable corresponding to the sum of the cubes of the velocities plus the square of the masses of all of the component molecules of the gas. However, this variable will not stand in any stable or coherent relationship to other upper-level variables that we are able to measure or characterize for the gas. Given some specified set of lower-level variables and laws or generalizations governing these, upper-level variables that are not just multiply realized by those lower-level variables but stand in stable, realization-independent relationships to one another are often rare and difficult to find. Indeed, in his influential text, Callen, 1985 claims that the familiar thermodynamic variables are the only ones that are constructable from (functions of) underlying statistical mechanical variables that stand in stable relationships. So finding upper-level variables that are multiply realized by lower-level variables *and* that are related by a generalization that is realization-independent (or comes close to being realization-independent) can be a highly non-trivial achievement.[9] Interesting examples of multiple realizability involve generalizations that have this feature of realization-independence.

As a point of comparison consider an example due to Spirtes and Scheines, 2004. Suppose that high density cholesterol, *HDL*, has a beneficial effect on heart health *H* and low density cholesterol, *LDL*, has a deleterious effect. Define total cholesterol *TC* as the sum of *HDL* and *LDL*. Values of *TC* are thus multiply realized by sums of pairs of values, one from *HDL* and one from *LDL*. But *TC* does not have a stable, realization independent effect on *H*, since the impact of any value of *TC* on any particular occasion will depend on the mix of *HDL* and *LDL* that happens to realize that value on that occasion, something that will be different on different occasions. Intuitively, if our interest is in heart health, *TC* is at the "wrong" level[10] or at least

[9] I acknowledge that on this understanding there will be cases of MR that are uninteresting—for example, masses of different colors will realize the same relation involving gravitational force. I doubt however that any formal general characterization of MR (or conditional independence) will be able to fully distinguish the interesting from uninteresting cases. More context-specific information is required. "Interestingness" is not a formal notion. That said, it is natural to restrict the candidates for realizers to which it is worth paying attention to those variables in a relevant science. Putting aside color science itself, color distinctions are not important in physical theorizing, while differences in, say, mass and charge are. So when we find systems that differ in mass and charge behaving in the same way, we think of that as an interesting case of MR, but not so for systems that differ in color.

[10] In comments on this chapter (personal communication), Larry Shapiro remarks that he does not find it natural to regard *TC* as at a different level from *HDL* and *LDL*. Nothing much turns on this

it is a "bad" variable. The variables *HDL* and *LDL* are at a much better level for forming generalizations about heart health. Note also for future reference that on this analysis the problem with *TC* is *not* that it is a "disjunctive property" or not a genuine kind or that it combines "causally disparate" realizers. The problem is rather that *TC* does not have a uniform, realization independent effect on *other* variables of interest like *H*. As argued below, it is difficult to translate the notion of a disjunctive property into a framework in which causal relata are variables, but in any sense in which *TC* is disjunctive, average kinetic energy also appears to be disjunctive. What makes average kinetic energy a "good" variable is that it has a stable realization-independent relation with other variables—it is not a matter of whether it is disjunctive or is a legitimate "kind".[11]

14.5 Conditional Independence, Levels and Multiple Realizability

With this as background, we can now proceed to the characterization of conditional independence and its relationship to levels and MR. Suppose as before that we have two sets of variables $\{U_i\}$ $i = 1, 2, \ldots$ and $\{L_j\}$ $j = 1, 2 \ldots$ with values of the latter realizing values of the former. Let us say that a variable X (whether upper or lower-level) is unconditionally causally relevant to a second variable E if there are some changes in the values of X when produced by interventions that are associated with changes in the value of E. (Thus unconditional causal relevance is what is just captured by the interventionist criterion for causation mentioned earlier). Suppose as before that we have a variable U_i and lower-level variables $\{L_j\}$ the values of which realize values of *Ui.* Suppose in addition that U_i is unconditionally causally relevant to E and that its realizers in L_j are also unconditionally relevant to E. Then those L_j

issue but for what it is worth I find this levels claim natural because *TC* is a coarsening of {*HDL, LDL*}, values of the former contain less information than pairs of values of the latter and so on. But again my intention is not to try to capture everyone's intuitions about levels.

[11] Although I cannot address the underlying issues here in the detail that they deserve, it is worth noting, that worries about predicates etc. that are disjunctive or otherwise illegitimate goes back at least to Goodman, who introduced a contrast between those predicates that are projectible and those that are not. This contrast was then transmorgrified by Quine, Lewis and others into a distinction between "natural" properties and kinds and those that are not. Although departing from Goodman in other respects (in particular in no longer regarding projectible predicates as simply those we in fact project), these writers retained the idea that there was something about the predicates or properties that figure in candidate generalizations that determines whether these are "laws" or otherwise legitimate. However, science just doesn't work that way—the various scientific disciplines countenance laws and generalizations containing terms of arbitrary complexity and that combine other variables in complicated ways (tensors representing stresses, integrals of distributions that are everywhere zero except at a single point etc.). These are regarded as legitimate insofar as they allow for the formulation of stable generalizations. So the focus on legitimate versus illegitimate kinds/properties as the key to "lawful" generalizations goes wrong at the start.

are *conditionally irrelevant* to E (or *conditionally independent* of E) if conditional on the values of U_i, changes in the values of L_j that are consistent with those values of U_i make no difference to E. Here relevance and conditional irrelevance should be understood in terms of counterfactuals rather than as probabilistic or statistical relevance and irrelevance.[12] That is, we are to imagine that interventions on values of U_i and separate interventions on values of L_j will both change E, but that when U_i is set to some such value u^* via an intervention, independent interventions on the realizing values of L_j that realize u^* (and thus that are consistent with $U_i = u^*$) do not change E. Strict conditional independence requires that this be true for all values of U_i within some range of interest. However, this condition can be relaxed in various ways—see below. Informally, U_i "screens off" L_j from E, where this screening off relation is understood in terms of interventionist counterfactuals, rather than conditional statistical independence. To return to our earlier illustration, conditional on the setting of the temperature of a dilute gas to some value $T = t$, further variations in the individual kinetic energies of component molecules of the gas that are consistent with $T = t$ will have the same effect on various other thermodynamic variables E such as pressure.

This is my attempt to make the ideas about levels, lower-level details not mattering and multiple realization discussed earlier somewhat more precise. When a conditional independence relation of the sort above holds, we may think of the *relationship* between U_i and E as multiply realized, with lower-level details concerning the realization of U_i not mattering to that relationship in the sense that all the causal or nomological information that is relevant to E in L_j has been absorbed into U_i and this entitles us to use U_i rather than L_j in accounting for E.

This framework has a number of additional features that are worth noting. First, conditional independence is a three termed relation, involving not just an upper-level variable and its realizers but a third variable E—some target variable or effect or set of these that we want to explain. It is entirely possible for lower-level variables L_j to be irrelevant to variable E_1, conditional on upper-level variable U_i and yet for L_j to be relevant to some other variable E_2 conditional on U_i. For example, there are many potential explananda (e.g., facts about specific heats of gases) to which lower-level details about individual molecules are relevant and that are not captured or screened off by classical thermodynamic variables. In other words, conditional independence does not require that there be no differences among the realizers of U_i or that these realizers have exactly the same effects in all respects, but rather simply that there be uniformity of effect with respect to E (or perhaps some set of Es—e.g., other thermodynamic variables). I will say more about this below.

Second, note another feature of the conditional independence relation. As I have characterized it, under conditional independence *both* the lower-level realizers and the upper-level variables that they realize are unconditionally relevant to some target variable E. Thus our discussion so far does not imply that the lower-level variables cannot be used to explain E or that they provide a "worse" explanation of E than

[12] For more details on how this works, see Woodward, Forthcoming.

the upper-level variables. Rather what follows from conditional independence is simply that we don't need to advert to the lower-level variables to explain E, and that we can use the upper-level variables instead. This contrasts with the common tendency in the philosophical literature (e.g., in both Putnam and Fodor) to argue that upper-level explanations of upper-level explananda are (always?) "better" than their lower-level counterparts. I reject this general claim, although there are some subtleties here to which I will return below.

Third, the condition just described represents a kind of *ideal* of complete conditional independence. It can be relaxed in various ways. For example, one might think in terms of "effective" or approximate conditional independence, meaning by this that the departures from full conditional independence are "small", with little loss of information in using U_i rather than L_j.[13] Or it may be that although conditional independence of L_j from E given U_i does not hold for all values of L_j and U_i it holds for "almost all" such values or perhaps all within a certain large interval, where the interval describes commonly occurring boundary conditions and constraints. For example, effective conditional independence of molecular level variables for certain aspects of cellular behavior given certain cellular level variables may hold for many sets of variable values that obtain within a living cell, even if not for all such values—e.g., those that are incompatible with the continued life of the cell. Finally, when even approximate conditional independence fails, one can sometimes restore it by adding additional variables to candidate screening off variables—for example, perhaps variables that are at an intermediate or meso-level.

Although I lack the space to argue for this claim in detail, I believe that when these additional possibilities are taken into account, there are a number of realistic cases of conditional independence or near conditional independence. In any case, I emphasize that how widely conditional independence holds is an empirical matter and not something that can be decided from the armchair.[14]

14.6 Kinds: Natural and Otherwise

A very substantial part of the literature on multiple realizability and the "autonomy" of upper-level science generalizations, including classical discussions by Fodor

[13] See Ay and Polani, 2008 for a proposal about how to measure this information loss. I will add that although some philosophers may think that the introduction of "near" or "effective" conditional independence is a cheat, my view is that all known scientific laws and theories are merely "effective" and hold only under certain conditions. In countenancing effective conditional independence (rather than perfect conditional independence for all variable values), we are invoking a feature that is ubiquitous in science.

[14] For discussion of additional examples, see Woodward, Forthcoming. I have encountered a number of pronouncements (without further detail) by philosophers that real cases of conditional independence or even near conditional independence virtually never occur. I repeat that this is an empirical issue that requires attention to real examples.

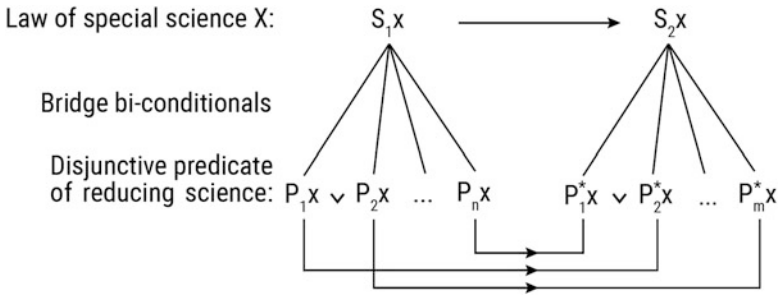

Fig. 14.1 Fodor's diagram. (Based on image (5) from Fodor, 1974: 109)

1974 and Kim 1993, is organized around notions of "kinds" and/or properties that are ascribed to kinds. This focus persists in more recent discussions (Polger and Shapiro, 2016). A background assumption to this discussion, explicit in writers like Fodor, 1974, is that most or all legitimate causal generalizations or laws (at least insofar as the special sciences contain these) relate "kinds" (often taken to be "natural kinds") or natural (legitimate, non-disjunctive or gerrymandered) properties associated with kinds: emeralds, ravens and copper are kinds, green is a natural property and so we have causal generalizations like "emeralds are green", "ravens are black", "copper conducts electricity" and so on. Generalizations that do not involve such kinds are thought to be bad candidates for special science generalizations. Multiple realizability is then conceptualized as a matter of different lower-kinds realizing the same upper-level kinds—e.g., human and Martian brains (assuming these to belong to different kinds) or different kinds of states of these realize some common psychological level kind such as pain. Upper-level causal generalizations link such multiply realized kinds to other upper-level, multiply realized kinds or properties—e.g., pain causes avoidance behavior (also assumed to be multiply realized).[15]

The iconic diagram above (Fig. 14.1) from Fodor, 1974, 103 captures the basic idea—notice that in this picture a number of lower-level different "laws" or generalizations, one for each of the kind predicates P_1, P_2, ... P_n underlie the single upper-level law.

Note that within this framework different underlying laws are required because laws are individuated in terms of the different kinds or properties whose behavior or effects they describe and the properties P_1, P_2 etc. are different. This contrasts with the account of MR in Sect. 14.5, in which in standard cases the same law or laws—e.g., those governing the statistical mechanics of molecules—describes the behavior of each of the realizers and in which the realizers are conceptualized in terms of values of variables (typically the same variables but with different combinations of

[15] I focus in what follows on kinds, but many of the points that follow also hold for "properties", as philosophers conceive of these. Property talk suffers from many of the same limitations as kind talk, in its failure to map on to a variable-based framework.

values). Given this Fodorian picture, the basic issue concerning MR concerns how (or whether or when) "different" underlying kinds realize the "same" upper-level kind, thus immediately embroiling us in questions about sameness of kinds and when a purported upper-level kind is a genuine, legitimate kind or illegitimately disjunctive.

It might seem that the kind-based formulation of MR differs only in minor ways from the formulation in Sect. 14.5. I think this is wrong. Despite its popularity, this kind-based formulation and the various background assumptions with which it is associated (e.g., that laws and causal generalizations in science are typically formulated in terms of kinds) have been a major source of confusion in discussions of MR—indeed they make it difficult to make sense of this notion if one wants to use it for the purposes described above. Or so I shall now argue.

Let me begin with the notion of a kind (or natural kind). It is true enough that various sciences recognize or make use of kinds—one thinks of chemical elements (or perhaps their various isotopes) as kinds, protons and electrons may be described as kinds of particles, neurobiologists recognize different kinds of neurons and so on. But many if not most serious candidates for laws or causal generalizations are not naturally regarded as describing kinds or their behavior. This is so for several different reasons. First, the notion of a kind, when applicable at all, is a binary notion—a substance is either of the kind copper or not. By contrast, as assumed in previous sections, typical laws relate quantitative variables or graded magnitudes like mass, charge, distance, force, time and so on. These are not kind terms on any reasonable understanding of that notion. Even when kind talk makes sense, as in connection with different kinds of atoms—hydrogen, helium and so on—the laws governing the behavior of these will be quantitative laws (the Schrödinger equation etc.) that provide a common treatment of all of the kinds in question, and the relevant variables will be mass, charge and so on, which will apply (with different values) to all of these kinds. That is, differences among these kinds are modeled in terms of the same set of equations involving a common set of variables applied to different initial and boundary conditions. There aren't separate laws or separate variables for the kinds hydrogen, helium and so on, contrary to what is assumed in Fodor's diagram.[16] A similar point holds for causal generalizations of the sort discovered in the special sciences—a causal generalization relating the social economic status of a person's parents and the educational level of the parents to that person's

[16] Fodor's assumption that each of the realizing kinds involves a separate law makes their common realization in a single upper-level law look more puzzling and difficult to explain than it actually is because it suggests that the realizers have nothing interesting in common that can be given a "lower-level" characterization—a conclusion that Fodor adopts. But it is often possible to explain why and to what extent lower-level details don't matter to an upper-level generalization. This is often possible in part because there is a single set of underlying laws for realizers or at least a common framework for their representation. For example, in connection with the explanation of uniformities in critical point behavior, the common representation of the various systems in terms of characteristics of their Hamiltonians is crucial to the renormalization group analysis that allows us to see why certain lower-level details don't matter (Batterman, 2021).

educational level or income relates quantitative variables rather than kind terms. Similarly, a computational model of human visual processing may make use of differential equations relating such quantitative variables as light intensities—again this is not a story about "kinds". Computational models of human causal cognition, as in Cheng, 1997 make use of variables that reflect patterns of correlations and structural equations—again nothing that looks kind-like.

As another illustration, even if one thinks of "pain" as a kind (itself a strange idea, given the many different non-binary dimensions along which pains can vary), its physiological underpinnings (the pain matrix) or realizers involve a large number of distinct neural structures[17] executing different computations, the actions of various neurotransmitters, nociceptors throughout the body and so on. Even on a particular occasion on which pain is experienced, these realizers will not belong to a single kind and whatever the correct physiological/neurobiological account of pain may be, it is not going to take the form of the identification of some single kind of "thing"—a kind of neuron, an anatomical region of the brain or even a single brain "state"—that realizes "pain". Nor is it likely to take the form of two or more such kinds—e.g., one characteristic of humans and others characteristic of other species, each of which realizes pain. This is not because pain can't be multiply realized but rather because thinking of such realization in terms of relations between neurobiological and psychological kinds is just inapt.[18]

These differences between generalizations formulated in terms of kinds and those formulated in terms of variables are related to an even more fundamental difference. Many/most laws and special science causal generalizations describe how changes or variations in quantities or magnitudes relate to changes or variations in other magnitudes. The fact that such laws or generalizations are formulated in terms of variables that can take different values is crucial to this. For example, the gravitational force law describes how variations in the masses of bodies or the distance between them are associated with variations in the gravitational force between them. In this way the law describes how the gravitational force depends on these other factors. Even a low-level garden variety causal generalization like

(1) Aspirin ingestion causes headache relief

has a broadly similar structure—it is naturally understood as claiming the existence of a dependency relation between each of the values of two binary variables: a

[17] For example, pain commonly involves the activation of brain stem, somato-sensory cortex and more frontal structures such as insula and anterior cingulate cortex, but with variations in each area depending on the nature of the pain involved.

[18] Another difference: generalizations involving kinds typically relate just two properties, the kind and some characteristic feature (copper melts at 1983 degrees etc.). By contrast typical laws and special science causal generalizations relate more than two properties, with several distinct magnitudes or variables occurring in their "antecedents". For example, according to Stokes' law the force F on a sphere of radius a moving through a fluid of viscosity η at speed v is given by: $F = 6\pi a\eta v$. Some contortion is required to reconstrue this as a claim about a property attaching to a single "kind".

change from a situation in which you ingest aspirin to one in which you do not or vice-versa is associated with a change in whether (or perhaps the probability of whether) you have headache relief. Assuming that you have a headache, there is a dependency relation of some kind between whether or not you ingest aspirin and whether your headache goes away soon: you will be more likely to have relief soon if you ingest aspirin than if you do not. We might capture this by thinking of (1) as relating two variables each of which is two-valued: variable *A* the values of which correspond to whether one ingests aspirin or not and a variable *R* corresponding to whether one experiences relief or not. A similar point holds for the various examples of special science causal generalizations noted in previous paragraphs.

The fact that most laws and causal generalizations, whether in fundamental or special sciences, have this dependency-relating feature marks a fundamental difference with the generalizations about kinds that figure in philosophical discussions, including those that center on multiple realizability. In most cases such kind generalizations are not naturally interpretable as describing dependency relations. Instead, at least as understood by philosophers, they purport to identify a condition C (membership in some kind K) that is (supposedly) "nomologically sufficient" for some outcome but without telling us anything about what would happen if C were to be different or to change—a feature which is crucial for dependency information. For example, even if the kind generalization

(2) All emeralds are green

is true, (2) is not readily understood as telling us that whether or not something is green "depends on" whether or not it is an emerald. (2) claims that being an emerald is sufficient for being green but is silent on the conditions under which non-greenness occurs. (We certainly cannot interpret it as claiming that all non-emeralds are non-green). In this respect it is quite different from what is conveyed by (1).

The idea that causal generalizations and many laws describe dependency relations is implicit in the interventionist treatment of causation to which I appealed earlier. It is also reflected in the common idea that causes are difference-makers for their effects. This is why I claimed that absence of this feature in kind generalizations like (2) marks a deep difference between them and, alternatively, causal generalizations and many laws. Moreover, on the analysis in Sections 14.4 and 14.5 both conditional independence and MR have to do with patterns of dependency and independence relations. If anything along the lines of this analysis is correct, it will not be surprising if discussions of MR that are framed in terms of kind-generalizations that do not express such dependency information will miss much that is important and are likely to lead to problems that are artifacts of this framing. In the following section, I show in more detail that this is the case.

14.7 Realization and Disjunctive Kinds and Properties

Consider a kind-based framework for understanding MR and suppose that pain is our candidate upper-level cause variable (the associated effect being something like avoidance behavior) and that pain is realized in humans by some carbon-based kind C and by some silicon-based kind S in Martians.[19] How might we think about this in terms of a variable-based framework? An initial thought is that the kinds C and S might be interpreted in terms of variables X and Y (where we are assuming that X and Y are distinct), with X taking the value 1 or 0 depending on whether C is present and Y taking the value 1 or 0 depending on whether S is present. However, this won't work, since the presence of C in any individual excludes the presence of S in that individual and conversely—the same individual can't belong to both the kinds S and C. This means that various combinations of values of X and Y (such as $X = 1, Y = 1$) are not compossible and, according to the criterion defended above, (Sect. 14.3) that X and Y are not distinct variables. The relation between C and S (and X and Y) looks more like the relation we expect to obtain between values of the same variable (which as we have seen do exclude each other) than the relation we would expect if X and Y correspond to distinct variables. Recognizing this, we might consider employing a single variable, call it Z, with the presence of S and the presence of C corresponding to different values of that variable. (This captures the incompatibility between S and C). However, this immediately introduces another complication. Since it presumably must be possible for a creature not to be in pain, Z will need to have at least one additional possible value—call it z_3—corresponding to whatever realizes the absence of pain in humans and Martians.[20] (If the presence of S or the presence of C are the only possible ways in which pain can occur, then the absence of pain will correspond to situations in which neither S nor C is present. In any realistic case, we will likely need more than one value for absence of pain, but again let's put that aside). We need this additional value since any generalization in which pain figures as a cause will need to specify what happens when pain does not occur (or at the lower level what happens when the realizer of pain in humans/Martians does not occur) as well as what happens when pain or its realizer does occur, assuming, as we are, that causal generalizations describe difference-making or dependency relations. We should also note, as more evidence for the problematic character of a kind-based framework for thinking about this example, that we now seem led to thinking about the absence of pain as itself a kind or collection of kinds (alongside pain as a kind) and (apparently) the absence of C and the absence of S as a kind or kinds that realize the absence of the kind pain. Needless to say one does not usually think of absences (at least of this sort) as kinds. However, this seems to be required if we are thinking in terms of upper-level generalizations and associated lower-level

[19] The idea that these pain realizers are "kinds" is already absurd for reasons described above as is the excursion into science fictional Martians, but I will ignore this in what follows.

[20] Recall that we have assumed that the function describing the realization relation is surjective, which requires that there must be a value realizing absence of pain.

realizing generalizations that are formulated in terms of kinds and if we want to capture the dependency aspect of these generalizations.

These considerations by themselves suggest that the kind-based framework for thinking about MR does not work smoothly when we think of causal generalizations as expressing dependency relations. Moreover, an additional difficulty surfaces when we turn to an issue that has been central motivation for the superiority of upper-level generalizations within a kind-based framework. This concerns so-called disjunctive properties or laws with disjunctive antecedents. Suppose that pain is multiply realized in the manner described above. Assume that legitimate laws or generalizations must be formulated in terms of kinds and consider a law or generalization formulated in terms of the pain realizers kinds C and S—e.g.,

(3) If C or S, then pain behavior B.

According to Putnam and Fodor, (3) is not a legitimate law because it involves a disjunctive predicate or antecedent: C or S is not a natural kind. By contrast (they argue) the upper-level generalization

(4) If pain, then behavior B

is a candidate for a legitimate law because "pain" is non-disjunctive and a legitimate kind or property. This is claimed to show that explanations in terms of (4) are superior to explanations in terms of (3), thus vindicating the use of the upper level (4).

I will say more about this and related arguments below, but let us first see what it might look like in the variable-based framework that I have been advocating. I claim that within this framework the argument just described cannot be coherently formulated. (Of course I think this is an objection to the argument, not to my framework. It is an advantage of my framework that the argument cannot be formulated). Within the variable-based framework the realizer for pain or its absence will be something like values of the variable Z considered above (or more plausibly values from some vastly more complicated set of variables, but again let's put that aside). Recall that Z is a variable that can take three values, one corresponding to the presence of C, one corresponding to the presence of S, and one corresponding to their absence. Is Z (or does it correspond to something) "disjunctive"? This seems to be a very inappropriate description. Instead, Z is just a variable that can take three values. The fact that Z can take more than one value does not make it disjunctive or if it does, all variables are disjunctive, since all variables must take more than one value. For comparison consider the variable "mass". Should we consider mass a hugely disjunctive variable on the grounds that it can take any real positive value? Even if you are tempted to say, "yes", it seems obviously misguided to go on to claim (à la Fodor) that generalizations in which mass occurs are illegitimate or non-lawlike on the grounds that mass is a disjunctive property or that "mass" does not correspond to a natural kind. There may well be something inapt or defective about Z as a variable but if so, this is not a matter of its being disjunctive. So, contrary to what many have supposed, one can't argue for the superiority of (4) over (3) on the grounds just described.

Indeed, the whole notion of a disjunctive variable (as opposed to a disjunctive property) seems highly problematic.[21] To drive the point home, consider the natural representation of an ordinary "or" gate—if anything counts as disjunctive it is surely an "or" gate. Suppose that A and B are variables representing inputs to such a device and C the output, and that these are related by the following structural equation, with each variable having possible values 1 and 0 corresponding to true, false.

(5) $C = \max (A, B)$

The functional relation between A, B as inputs and C as output behaves like "or" or "disjunction" but is a "disjunctive variable" or even a "disjunctive property" present in this example? No. A functional relation involving three variables that behaves like a disjunction is present but that is not the same thing as a disjunctive variable. In the above interpretation, A and B are ordinary distinct two-valued variables, with their distinctness shown by the fact that all the combinations of their values are possible, in the state space sense described above. Distinctness is also exhibited in the fact that in using this representation we are claiming that, for example, if $B = 1$, we can intervene separately to set A to either 1 or 0 while leaving the value of B undisturbed.[22]

I conclude from this that the whole notion of a disjunctive property or kind and associated claims about the status of generalizations containing these rests on misunderstandings about the structure of the causal generalizations that figure in science—failures to recognize that such generalizations relate variables and describe dependency relations.[23]

With this as background let return to some of the examples discussed in previous sections. Consider the case of total cholesterol *TC* which is the sum of *HDL* and *LDL*, with pairs of values of the latter realizing values of *TC*. Some may be tempted

[21] Just to be clear: I'm not claiming that the notion of a disjunctive property or predicate is incoherent. Obviously I can talk about whether Pa v Qa holds for individual a and so on. I am claiming that there is nothing that straightforwardly corresponds to this when we talk in terms of variables. Also in this connection, let me add that I see the discussion in Sober (1999) as consistent with, but somewhat orthogonal to mine. In contrast to my discussion Sober focuses on the notion of a disjunctive property.

[22] As an alternative we might consider modeling this situation in terms of a single variable D which takes the values 1 iff $A = 1$ or $B = 1$ and 0 otherwise. Then the corresponding structural equation would be

$C = D$.

This has the disadvantage that we have no longer represented the fact that we can set the values of A and B separately but putting this aside, there still does not seem to be any sense in which D is a "disjunctive" predicate or variable—again it is just an ordinary variable that takes two values.

[23] Although exploration of this point must be beyond the scope of this essay, the idea that generalizations in science have primarily to do with kinds rather than quantitative variables seems to hark back to a broadly Aristotelian picture of science and an accompanying metaphysics: generalizations in science describe tendencies or dispositions that are inherent in kinds of objects. This simply doesn't fit very well with most of modern science.

to say, in the spirit of Fodor et al., that *TC* is not a "good" variable because (i) it is disjunctive and/or (ii) because it combines values from each of two distinct kinds, *HDL* and *LDL* and that is illegitimate, perhaps showing that *TC* is not a proper kind. But, as argued above, there does not seem to be any clear sense in which *TC* is disjunctive. As far as (ii) goes consider temperature understood as the result of averaging the kinetic energy of each of the component molecules of the gas, with the kinetic energy of each corresponding to a distinct variable. This combining of values of these distinct variables does not make average kinetic energy an illegitimate variable. For that matter consider kinetic energy itself (predicated, say, of an individual molecule) which combines values of a mass variable, a distance variable and a time variable. Presumably no one thinks that makes kinetic energy an illegitimate variable, unsuitable for service in laws. Again we see how a kind-based framework, which encourages us to ask such questions as whether kinetic energy is a legitimate kind and how this relates to the fact that it is "built" out of such disparate factors as mass, distance etc., is highly unperspicuous.

Despite this I agree that *TC* is not a "good" variable for the purposes of explaining heart health. On my view, the reason why it is not a good variable is simply that it does not seem to figure in stable, intervention supporting upper-level generalizations concerning heart health (or any other effects that we know of). By contrast, temperature does figure in such generalizations. In neither case does this have anything to do with the disjunctiveness or kindhood or lack of kindhood of the variables figuring in these generalizations.

14.8 Kim and Polger and Shapiro on Multiple Realizability

I turn now to some comments on two other influential discussions of MR, both conducted in a kind-based framework and hence subject to its infirmities. Kim, 1993 considers cases in which two or more lower-level kinds, L_1 and L_2 are claimed to multiply realize a single upper-level kind U. He argues that the assumption that L_1 and L_2 are genuinely distinct (as is required by the usual understanding of MR) is in considerable tension with the claim that U is a single, non-disjunctive kind (which it must be if, according to the Putnam/Fodor argument, it figures in legitimate upper-level generalizations). Kim asks if, as is commonly assumed, there are bridge laws connecting L_1 or L_2 to U and these are in some sense "necessary", why U doesn't inherit the disjunctiveness of L_1 or L_2. Seeing that U has two completely distinct realizers, why don't we conclude that U should be split into two distinct kinds, one corresponding to L_1 and the other to L_2? One of Kim's examples is the candidate kind jade which Kim thinks of as having two different realizers—jadeite and nephrite. He suggests that this difference in realizers shows that jade is not a single unified kind. In effect, the puzzle that Kim raises is how an upper-level kind can be both non-disjunctive (required for it to be a genuine kind) and yet be realized by distinct kinds, as MR requires.

A somewhat similar line of argument is advanced more recently by Polger and Shapiro (2016). Their "official recipe" (p. 67) of what is required for MR is complex but their basic strategy is to exploit the same-but-different tension described above. Like Kim, they formulate MR in terms of claims about kinds (or relations between what they call "taxonomic schemes" which basically have to do with kinds). Surveying a number of putative examples of MR, Polger and Shapiro argue that in many, perhaps most such cases, either the candidate realizing kinds are not sufficiently or relevantly distinct (so that there is not more than one realizing kind) and/or that the candidate upper-level realized kind is not really a single kind because instances of it are relevantly different. For example, they argue (pp. 44ff) that the octopus eye and the human eye are not distinct realizations of the same kind because, despite obvious anatomical differences, they are both camera eyes—they operate by using an iris to control the light focused on a photoreceptive surface. In connection with well-known experiments (Von Melchner et al. 2000) involving ferrets who learn to see with their auditory cortex, they argue that because the "rewired" ferrets perform differently on certain visual discrimination tasks from normal ferrets, the two sorts of ferrets do not realize the same kind. But they also seem to suggest that the two sorts of ferrets (or the relevant parts of their brains) may not really belong to different kinds since the visual cortices of the former and the auditory cortices of the latter display various similarities such as columnar organization (p. 96). The upshot of this strategy is that there are far fewer cases of genuine MR than many have supposed—either the realizers are not "really" different or, to the extent they are different, this supports the conclusion that they do not realize the same upper-level kind.

I have several, interrelated comments about such arguments. First, when we attempt to formulate them in a variable-based rather than a kind-based framework (or a framework that involves properties in the sense philosophers have in mind), they seem to have far less intuitive pull. If I claim that different combinations of values for the kinetic energies for the individual molecules in a gas can realize the same value of temperature $T = t$, there does not seem to be any very compelling basis for claiming that this suggests that the value $T = t$ should be "split" into distinct values, one for each combination of kinetic energies that realize t. At the same time, there also does not seem to be any good basis for claiming that the different realizing combinations of kinetic energies of the component molecules are not relevantly distinct. Kinetic energy is a paradigmatic physical variable and particles with different kinetic energies (and collections of these) have different properties and effects which in some cases are readily detectable. Similar points hold for the other examples of multiple realization that have been discussed in the literature. As discussed by Batterman in a number of books and papers (e.g., 2021), gases of different kinds as well as ferromagnets exhibit similar behavior, characterized by sameness of the exponent β in the relation $\psi = (T - T_c/T_c)^{\beta}$ (where ψ is the order parameter) near their critical points—a clear case of multiple realization of that relationship. Of course gases and ferromagnets differ in many other respects but it seems unmotivated to claim that because of these differences, the sameness in their behavior near their critical points is not genuine or that the

relation describing critical point behavior must be "split" into two relations, one governing gases and the other governing ferromagnets.

Part of what is going on here is that claims about sameness, similarity or differences among kinds are often vague and contestable and certainly relative to what is taken to be the relevant dimension for assessing sameness or difference. An object can belong to many different "kinds" and there may be no objective basis for picking out just one of these as *the* basis for judgments of sameness and difference. If the relevant kind is "camera eye", then human and octopus eyes belong to the same kind. Given some other basis for classification (nature of cells involved, positioning of retina) human and octopus eyes will belong to different kinds. If the relevant kind is something like "visual processing with features X" then normal and rewired ferrets may exhibit that kind for some such features but not for others. If we impose a criterion for sameness or same kindness of visual processing, requiring capacities for visual discrimination not possessed by rewired ferrets, then of course the visual processing of the two sorts of ferrets will differ in "kind". Kim's arguments and those of Polger and Shapiro exploit these features of kind judgments via a kind of "heads I win, tails you lose" strategy. Given a putative case of MR, it will virtually always be possible to find some respects of difference between the different realizers of the candidate upper-level realizing kind. If there is no such difference, how can the realizers be different? This can then be used to support the claim that there is no single upper-level kind which they realize. At the same time, as remarked in footnote 16, it is far from clear that there any real cases of MR in which the lower-level realizing kinds have literally nothing relevant in common. This fact then can be used to support the claim that these realizing kinds are not really different, again showing the absence of MR on conceptions like those of Kim and Polger and Shapiro. One of the attractions of the variable-based framework for thinking about MR is that it does not rely on such arbitrary judgments about sameness or difference in kind membership (or property similarity). When understood along the lines described in Sect. 14.5, claims that various upper-level variables take values that are realized by values of lower-level variables and that the upper-level variables stand in stable relationships of various sorts are straightforward to assess and do not require problematic judgments about extent of sameness. For example, the claim that two gases can be at the same temperature and conform to the same upper-level thermodynamic generalizations despite differing in molecular details does not require arbitrary judgments.

I noted above Kim's (and Polger's and Shapiro's) argument that, to the extent that lower-level realizers of an upper-level kind are genuinely different, this seems to support the conclusion that the upper-level kind is not a single, unitary kind. The variable-based account of MR in Sect. 14.5 does not license inferences of this sort. One reason for this is that this account is effect or explanandum-relative: the question is always whether, for some set of systems, if the candidate cause X (and perhaps other variables) characterizing those systems is multiply realized in such and such a way and the candidate effect/explanandum variable Y is multiply realized in such and such a way, there is a stable intervention-supporting relation between X and Y. This does not require that there be such a stable relationship between

X (or other upper-level variables) and all various other distinct variables $Y^* \neq Y$. In other words it is consistent with the systems for which this relation between X and Y holds being different in various *other* ways despite conforming to the same X-Y relationship. As we have seen ferromagnets and gases conform to the same upper-level relationship concerning behavior near their critical points but of course ferromagnets and gases differ in many other respects and many of these differences can only be explained in terms of "lower-level" information. Within the framework that I advocate, these lower-level differences are not an objection to the claim of multiple realization.

I also noted above that the treatment of MR advocated by Kim and Shapiro and Polger threatens to define MR out of existence by imposing demands that are not simultaneously satisfiable—the realizers need to be different kinds, but this is then treated as reason for regarding the realized upper-level kind as not a single kind. This consequence seems undesirable and is avoided by the framework I advocate. I think, however, that there is more that can be said. Suppose one thinks that Polger and Shapiro have the "right" characterization of MR and as a consequence it is at least rare and that many of the cases commonly thought to involve MR do not. By itself this conclusion is not very satisfying—it still remains the case that in a number of cases lower-level details do not matter (that is, conditionally) for the formulation of stable upper-level generalizations or do not matter very much. One would like (at least) to have a framework for characterizing what is going on when this happens and, to the extent this is possible, also some understanding of how and when such cases can occur. Characterizing MR in such a way that such cases are not instances of MR does not give us any insight into these questions.

My remarks in Sect. 14.5 are intended as one possible proposal about how to understand "details don't matter" claims, a proposal that I think helps to make sense of such claims. However my proposal that such claims can be understood as conditional independence claims does not do much to explain *why* or under what generic circumstances conditional independence occurs. This last issue is a complex one, in part because it is by no means clear just what would count as the sort of explanation we are looking for. For reasons of space and competence I will not try to address this question in any detail but I will remind readers of some previous remarks that gesture in directions where I think explanations of the sort desired may be found. The relevant considerations are disparate: they include facts about the relative magnitudes and patterns of distance dependence characterizing various forces (e.g., explaining why nuclear forces or gravity don't matter for certain phenomena), the existence of large differences in the time scales at which various effects occur which may allow the effects of some slow-varying processes to be summarized as constants (thus allowing considerable dimension reduction), the presence of large-scale constraints or boundary conditions that reduce the effective degrees of freedom characterizing lower-level processes, the presence of symmetries and scale invariances that also reduce degrees of freedom, families of dynamics characterized by large basins of attractions so that systems that differ in detail may nonetheless flow to very similar outcomes, the presence of selective constraints that funnel disparate systems to the same outcomes and much

else besides. Techniques like the use of the renormalization group to explain the kind of MR involved in critical point phenomena exploit many of these considerations.

Polger and Shapiro's preferred alternative to MR is an identity theory of some kind. Without delving into the details of such theories, let me just note that even if otherwise defensible, they seem to provide little insight either into how details-don't-matter claims should be understood or when and why we should expect them to hold. That is, merely asserting that some upper-level property or state is identical with some lower-level state tells us little about why just these features of the lower-level state (the ones involved in the identity claim) are the ones that matter for upper-level behavior and why we are justified in ignoring other lower-level features. Moreover, conditional independence claims are, as we have seen, relativized to a target set of effects or explananda, allowing for the possibility that lower-level details do not matter for effect E but may matter for some distinct effect E^*. By contrast, it is arguable that identity claims cannot involve such relativization: if lower-level L is identical with upper-level U then whatever effects L has (including lower-level effects) must be effects of U as well—any lower-level details encoded in L must be encoded in U as well if the identity is genuine. Put the other way around, if L causes or explains some E^*, then if U is identical with L, U must cause or explain E^* as well. This seems in tension with the idea that the upper-level theory drops details captured by the lower-level theory—details that may be relevant to some explananda but not to those in which the upper-level theory is interested.[24] Relatedly, as noted above, identity claims seem to require that the upper-level property or magnitude have the same dimensionality as the lower-level property with which it is identified (magnitudes with different dimensionality can't be identical), thus making it difficult to express the role of considerations having to do with reduction in dimensions and degrees of freedom that successful upper-level theorizing often requires. Again this puts pressure on the idea that the right way to understand relations between lower and upper-level theories is usually or always in terms of identities.

[24] It may seem that this difficulty can be avoided simply by finding some construction fully characterized in terms of the lower-level theory that omits the details that don't matter and then identifying that construction with the upper-level property, as when, according to some, temperature in a gas is "identified" with its mean kinetic energy. Although I lack space for detailed discussion, there are a number of problems with this suggestion. Although I agree that the proposed identities can be thought of as expressing or asserting that various lower-level details don't matter, I observe again that they don't give us any insight into why we are entitled to ignore these details. That is, the identities themselves don't give us an argument of the kind that we considered earlier showing that details of nucleon structure don't matter to chemical behavior. Second, the idea that the "bottom half" of the identity claim can be formulated just in terms of the resources of the lower-level theory often turns out to be mistaken. Even in the case of the temperature of a gas, the identification needs to be with mean kinetic energy of the gas *at equilibrium*, the latter being an "upper-level" notion. In other cases, far more upper or meso-level information (and not just information from the lower-level theory) is required to formulate any candidate identity claim. Finally, if what really matters is extracting from the lower-level theory the information that matters for upper-level behavior and representing it in the upper-level framework, it is unclear why such information extraction always has to involve discovering identities between lower and upper.

To this we may add that in many cases, although there may be complex relationships between upper and lower levels, with the latter constraining the former in various ways and some of the information in the lower level being relevant to the upper level, the variables and generalizations involved may live in sufficiently different conceptual spaces and may be sufficiently mismatched that identity talk seems like the wrong notion to capture such relations. Consider the Young's modulus for a solid material which measures the relationship between its tensile stress and its axial deformation and is an upper-level parameter. Its value is a kind of summary of very complicated facts about the molecular and meso-level structure of the material. Yet it is not clear that it makes much sense to look for some lower-level property or structure of the material to *identify* with its Young's modulus. For one thing, Young's modulus is only defined at the level of upper-level behavior. The way to understand the value of the modulus is that it throws away a huge amount of lower-level information that turns out not to be relevant to various upper-level behaviors that we want to explain while retaining what is relevant. It is not clear why it helps (or adds anything) to think about this in terms of an identity.

As a very different example, consider the relationship between a psychological defined process such as memory retrieval, a model of the computations that underlie this and the neural "hardware" in which these computations may be implemented, as in Rolls, 2021. To the best of my knowledge, neuroscientists rarely if ever use identity talk to describe such relations—and not of course because they are dualists. The language that is used instead invokes "neural substrates" that "encode" or "implement" computations, computations that show "how memory retrieval is accomplished" (Rolls, 2021, pp. 260ff) and so on. Neuroscientists don't seem to claim that memory retrieval is identical to (or just *is*) some underlying computation or that the computation can be identified with whatever implements it. Again, there is nothing mysterious about this—there can be all sorts of non-causal relationships and constraints of various sorts between variables and descriptions at different levels, without identity being the most appropriate way to characterize such relationships. From a scientific point of view, what matters is what information in lower-level theories needs to be passed to upper-level theories for them to do whatever explanatory job they are designed for, what constraints propagate from the former to the latter, and related questions. Answers to such questions need not involve claims about identifications, although perhaps they sometimes do.

14.9 Autonomy and the Explanatory Status of Upper-Level Generalizations

In this final section I want to return briefly to some issues about the explanatory status of upper-level generalizations and their "autonomy". As noted above, a number of philosophers claim that explanations framed in terms of such explanations are *superior* to those framed in terms of lower-level theories, at least if we abstract away

from supposed "pragmatic considerations" such as inability to actually construct the lower-level explanations. (Or at least they proceed as though the goal is to demonstrate such superiority—cf. Weslake, 2010, Franklin-Hall, 2016). While there is a defensible thought behind such claims (discussed below) many of arguments supporting superiority are unconvincing. I have already addressed and rejected one such argument—that lower-level explanations or the lower-level properties or kinds which figure in such explanations are "disjunctive" or otherwise unnatural and inferior for this reason to upper-level explanations. Another common argument is that upper-level explanations are more general than lower-level explanations and superior for that reason. For example, Putnam suggests that an upper-level explanation that appeals to the dimensions and rigidity of a peg to explain why it does not fit into a hole is superior to a lower-level explanation because the former will apply to pegs of different material and other characteristics as long as these share the dimensions and rigidity of the original peg. Fodor holds that psychological level explanations are superior to lower-level neurobiological explanations because the former would be applicable to non-human creatures that are psychologically similar to us while the latter would not.

When advanced in the forms just described, I think such arguments misfire. It is true that there is a dimension of generality (direct applicability to other pegs) according to which the upper-level peg explanation is more general than its lower-level counterpart. However, if as Putnam seems to assume, the lower-level explanation appeals to laws from microphysics of some sort (perhaps the quantum mechanics of solids) these laws will be highly general and will apply (albeit with different initial and boundary conditions etc.) to many other systems besides pegs. So to the extent that generality is an explanatory virtue, it does not unambiguously argue in favor of the upper-level peg explanation. In addition, it is far from clear that the mere fact that an explanation applies to a range of different systems shows anything about its goodness. Suppose that a proposed psychological or computational level explanation of some aspect of the human visual system can also be shown to be applicable to some other animals or to silicon-based systems that we are able to construct. Why should we suppose that improves the explanatory credentials of the account as applied to humans? Skepticism seems even more appropriate when, as with Fodor, the other systems are claimed to be merely possible (perhaps merely "metaphysically" rather than causally possible) rather than actual.[25]

That said, there is something in the vicinity of the arguments of Putnam and others that is defensible. The basic point is that the "upper-level" explanations can explain or help to explain explananda that the lower-level explanations cannot. (Note that this is different from the claim that the upper-level explanations explain the same explananda that the lower-level explanations attempt to explain, only better, a claim that I have argued is difficult to support). This basic idea has been developed

[25] Similar objections apply to the claim that upper-level explanations are superior because they are more unifying.

very clearly by Batterman in a series of books and papers (see e.g., Batterman, 2021), although the version that I will present may differ from his in some respects.

Consider again (the science fictionish) example of a derivation D of the overall state, characterized in terms of thermodynamic variables like temperature, of a gas from low-level information about the state of each component molecule at time t and the laws governing their interaction. Even if such a derivation were possible, there is much that (considered by itself) it would not tell us. For example, it would not tell us which alternative states of the component molecules would lead to the same values for these thermodynamic variables and which would lead to different values. Put differently D will not tell us about conditional dependency relationships about the matters just described. An upper-level explanation that appeals to thermodynamic information will provide such information. Note that this can be true even if we are unwilling to say of some particular sample of gas, that the thermodynamic explanation of its behavior is "better" than one that appeals to D. To this we may add that D will also not give us any insight into why gases that differ in the states of their individual constituent molecules nonetheless exhibit the same upper-level behavior—an explanandum which we need an upper-level characterization to even formulate.

The remarks in the previous paragraph emphasize the idea that upper-level explanations can provide information that a particular individual lower-level explanation like D does not. I think, however, that for many candidate upper-level explanations, particularly in the special sciences, this consideration is somewhat beside the point. Assume, as I think that we should, that an explanation requires something like the actual exhibition of a specific dependency relation between explanans and explanandum, so that simply asserting that E depends in some way on C or that E is derivable from C without further details, is, even if true, not an explanation. (Asserting that the current state of the stock market depends on unspecified underlying quantum mechanical facts is no explanation of stock market behavior). Then in many cases we are in no position to exhibit explanations of the behavior systems characterized in terms of upper-level variables by appeal to lower-level theories of the sort found in, say, fundamental physics. Moreover, if the argument in previous sections is correct, in at least some cases, we don't *need* to do this, since there are upper-level variables and theories which capture, via conditional independence, the information required for the exhibition and construction of explanations of upper-level behavior. Given that this is the case and the unavailability of the lower-level explanations, it is unnecessary to argue in addition for the superiority of the upper-level explanations.

What about the autonomy of upper-level generalizations and the special sciences in which they figure? On the account provided above, such autonomy, to the extent it exists, is always relative, for the same reasons that conditional independence is relative. A candidate upper-level generalization G, relating, say, U to E is autonomous with respect to some lower-level theory containing generalizations relating lower-level variables L, to the extent that G is a true, stable intervention-supporting relationship—that is to the extent that L is conditionally independent of E, conditional on U. Autonomy fails to the extent that we have to take into

account additional information from *L* beyond what is in *U* to formulate such a relationship about *E*. Thus thermodynamics is autonomous to the extent that we can explain thermodynamic phenomena just by reference to thermodynamic variables; autonomy fails to the extent that we have to employ information from outside of thermodynamics (e.g., from classical statistical mechanics or quantum mechanics) to explain various phenomena. Similarly, psychology is autonomous to the extent that we don't have to advert to neurobiology or biochemistry to explain phenomena that we think of as psychological. On this conception, autonomy is a matter of degree in several respects—for example, psychology may be autonomous with respect to some phenomena and not others and, as with conditional independence, autonomy may hold with respect to some range of variables and not others. A specified set of lower-level facts may limit upper-level possibilities in some respects but not others.[26]

Putnam and Fodor seem to think of the autonomy of the upper-level sciences in a very different way—they think that autonomy requires that the generalizations of the upper-level science be unexplainable, even in principle, from the lower-level science. But whether there are psychological generalizations satisfying conditional independence with respect to neurobiology and whether those psychological generalizations are unexplainable in terms of neurobiology are very different matters. I see no reason to tie autonomy claims to strong and often implausible claims about inexplicability.[27]

References

Alon, U. (2007). *An introduction to systems biology: Design principles of biological circuits*. Chapman and Hall.

[26] In real-life cases, we are very much interested in the relation between a set of lower-level facts that are less than the full supervenience base for certain upper-level facts and those upper-level facts. In such cases, the lower-level facts can constrain the possibilities regarding the upper-level facts without determining them. For example, facts about neural architecture and connectivity may constrain the kinds of computations that a neural system can perform without uniquely determining which computations are performed—instead, the supervenience base for these may include many additional facts about the organism and its environment.

[27] At one point, Fodor identifies the claim that psychology is autonomous with the claim that it cannot be reduced to neurobiology where by reduction he has in mind the standard Nagelian picture with type-type identities. But if the generalizations of neurobiology and the claimed identities are correct, then the generalizations of psychology derivable from these are also correct—their formulation does not require supplementation with neurobiological information. So from the perspective of conditional independence it follows, contra Fodor, that psychology would be autonomous in the sense I have described. Fodor is thus working with a very different notion of autonomy. Of course, many of us agree with Fodor that psychology is not reducible in the Nagelian sense to neurobiology but also think that psychology fails to be autonomous (in the conditional independence sense) in important respects.

Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems, 11*, 17–41.

Batterman, R. (2021). *A middle way: A non-fundamental approach to many-body physics*. Oxford University Press.

Callen, H. (1985). *Thermodynamics and an introduction to thermostatics*. Wiley.

Cheng, P. (1997). From covariation to causation: A causal power theory. *Psychological Review, 104*, 367–405.

Craver, C., & Bechtel, W. (2007). Top-down causation without top-down causes*, 22*, 547–563.

Ellis, G. (2016). *How can physics underlie the mind? Top-down causation in the human context*. Springer.

Eronen, M. I. (2015). Levels of organization: A deflationary account. *Biology and Philosophy, 30*(1), 39–58.

Fodor, J. (1974). Special sciences (or: The disunity of science as a working hypothesis). *Synthese, 28*, 97–115.

Franklin-Hall, L. (2016). High-level explanation and the interventionist's 'variables problem'. *British Journal for the Philosophy of Science, 67*, 553–577.

Goldenfeld, N. & Kadanoff, L. (1999). Simple lessons from complexity. *Science, 284*, 87–89.

Humphreys, P. (1997). How properties emerge. *Philosophy of Science, 64*, 1–17.

Kim, J. (1993). Multiple realization and the metaphysics of reduction. In *Supevenience and mind* (pp. 309–335). Cambridge.

Lewis, D. (1986). *Philosophical papers* (Vol. II). Oxford University Press.

Polger, T., & Shapiro, L. (2016). *The multiple realization book*. Oxford University Press.

Putnam, H. (1975). Philosophy and our mental life. In *Mind, language and reality: Philosophical papers* (Vol. 2, pp. 291–303). Cambridge University Press.

Rolls, E. (2021). *Brain computing: What and how*. Oxford University Press.

Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of Science, 66*, 542–564.

Spirtes, P., & Scheines, R. (2004). Causal inference of ambiguous manipulations. *Philosophy of Science, 71*, 833–845.

Von Melchner, L., Pallas, S., & Sur, M. (2000). Visual behavior mediated by retinal projections directed toward the auditory pathway. *Nature, 404*, 871–876.

Weslake, B. (2010). Explanatory depth. *Philosophy of Science, 77*, 273–294.

Woodward, J. (2008). Mental causation and neural mechanisms. In Hohwy & Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation* (pp. 218–262). Oxford University Press.

Woodward, J. (Forthcoming). Downward causation defended. In J. Voosholz & M. Gabriel (Eds.), *Top-down causation and emergence*. Springer.